MONTE CARLO EVALUATION OF ERROR RATE ESTIMATORS IN DISCRIMINANT ANALYSIS UNDER 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. This paper is concerned with the problem of estimating the error rate in two-group discriminant analysis. Here, behaviour 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 sum squared error (SSE). 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, on overall, the Finite Mixture Balanced bootstrap estimator (FMB) proposed by Mangku (2007) is the best estimator.

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

1. Introduction

One of the problems in two-groups discriminant analysis 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.

There are three types error rates that have been frequently considered for study, namely: (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, 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, 1974a), *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. Although various criteria have been looked at in the past to evaluate and compare the error rate estimators, the most popular choice was the *mean squared error (MSE)* criterion. One of the earlier authors employing this criterion was McLachlan (1974a, b, c), though more recently Efron (1983), Snapinn and Knoke (1984, 1985), Ganeshanandam and Krzanowski (1990), and some others have also utilized *MSE* in their evaluations. The *MSE* relative to the actual error rate is defined as $MSE = \mathsf{E}(\hat{P} - AC)^2$ where $\hat{P}$ denotes the estimated error rate, AC is the *actual error rate*, and the expectation is taken with respect to all possible sets of simulated training samples.

The criterion that we use in this study is the *sum square error*, denoted by *SSE*, equals to k x *MSE*. Hence, this criterion has similar properties to that of the *MSE* criterion. Here k is the number of simulated training data. The consequence is that the smaller is the *SSE* the better is the error rate estimator.

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

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

and

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

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 *SSE* criterion was then computed as

$$SSE = \sum_{i=1}^{50} (\hat{P}_i - AC_i)^2,$$

where $\hat{P}_i$ 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.* Note also that, 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 *SSE* 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 *SSE* with respect to the significant experimental treatment effects.

The statistical computing software SAS was used to carry out the above analysis. The result 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 shown 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.

Consider first the main plot stratum of Table 1. This table shows that, only the main effects of the experimental factors f and ν, and the effect due to the interaction p x ν are highly significant. This means that the estimation of the actual error rates, on average, is heavily dependent on the *size of training samples*, the *inter-dependence* of the variables, and the *number of variables* in the data. The Mahalanobis squared distance factor Δ^2 also has some noticeable effect (significant at 6%) on the error rates.

Further, the split plot stratum shows that the effects of *METH* factor and its interaction with p, f, p x f, Δ^2, p x Δ^2, f x Δ^2, ν, p x ν and f x ν are all significant. This not only suggests that there are differences in *SSE* among some of the estimators, but also indicates that the comparisons among these estimators must be based on the experimental factors p, f, Δ^2, and ν. The influence of the factor f or the size of samples, on the error rate estimators in a sum-square-error sense, is much higher than that of the factors ν, p, and Δ^2. This argument is due to the corresponding F-ratios being 102.15, 37.60, 34.07, and 28.10, respectively.

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	0.0301	0.0301	0.05	0.8183
f	1	20.9479	20.9479	37.45	0.0001
$p * f$	1	0.0024	0.0024	0.00	0.9486
Δ^2	2	3.6752	1.8376	3.29	0.0528
$p * \Delta^2$	2	0.3081	0.1541	0.28	0.7614
$f * \Delta^2$	2	0.9287	0.4644	0.83	0.4467
ν	1	21.8065	21.8065	38.99	0.0001
$p * \nu$	1	4.5692	4.5692	8.17	0.0081
$f * \nu$	1	0.6516	0.6516	1.17	0.2900
$\Delta^2 * \nu$	2	0.2379	0.1190	0.21	0.8098
d	1	0.3401	0.3401	0.61	0.4423
$p * d$	1	0.2266	0.2266	0.41	0.5298
$f * d$	1	0.0329	0.0329	0.06	0.8101
$\Delta^2 * d$	2	0.6095	0.3048	0.54	0.5861
$\nu * d$	1	0.3338	0.3338	0.60	0.4465
ERROR	27	15.1007	0.5593		
Split plot					
METH	18	23.7213	1.3179	383.53	0.0001
METH $* p$	18	2.1073	0.1171	34.07	0.0001
METH $* f$	18	6.3178	0.3510	102.15	0.0001
METH $* p * f$	18	0.6583	0.0364	10.64	0.0001
METH $* \Delta^2$	36	3.4759	0.0966	28.10	0.0001
METH $* p * \Delta^2$	36	0.5016	0.0139	4.05	0.0001
METH $* f * \Delta^2$	36	0.8158	0.0227	6.60	0.0001
METH $* \nu$	18	2.3257	0.1292	37.60	0.0001
METH $* p * \nu$	18	0.6410	0.0356	10.36	0.0001
METH $* f * \nu$	18	0.1612	0.0090	2.61	0.0003
METH $* \Delta^2 * \nu$	36	0.1333	0.0037	1.08	0.3523
METH $* d$	18	0.0117	0.0006	0.19	0.9999
METH $* p * d$	18	0.0386	0.0021	0.62	0.8819
METH $* f * d$	18	0.0181	0.0010	0.29	0.9983
METH $* \Delta^2 * d$	36	0.0285	0.0008	0.23	1.0000
METH $* \nu * d$	18	0.0398	0.0022	0.64	0.8660
ERROR	1486	1.6699	0.0034		

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	p x f	Δ^2	ν	p x ν
OS		21.310[a] 0.0001[b]			39.840 0.0001	10.350 0.0034
M		30.82 0.0001			21.37 0.0001	
NS	13.570 0.0010	51.710 0.0001		14.170 0.0001	56.950 0.0001	15.690 0.0005
R	6.38 0.0177	109.52 0.0001		18.87 0.0001	48.42 0.0001	10.72 0.0029
U	8.18 0.0081	56.75 0.0001			21.82 0.0001	5.25 0.0300
$\bar{U}$	11.00 0.0026	54.76 0.0001	5.73 0.0239	4.11 0.0277	13.98 0.0009	
JK	6.18 0.0194	56.87 0.0001			29.39 0.0001	4.29 0.0480
ISE		23.29 0.0001			36.51 0.0001	7.35 0.0115
IME		23.82 0.0001			38.41 0.0001	7.70 0.0099
ISC		24.12 0.0001			37.76 0.0001	7.45 0.0110
IMC		24.17 0.0001			39.24 0.0001	7.69 0.0099
ISB		23.62 0.0001			38.73 0.0001	7.45 0.0110
IMB		23.88 0.0001			38.46 0.0001	7.85 0.0093
FSE		22.62 0.0001			36.01 0.0001	7.40 0.0113
FME		23.48 0.0001			38.16 0.0001	7.73 0.0098
FSC		23.39 0.0001			37.28 0.0001	7.50 0.0108
FMC		23.80 0.0001			39.02 0.0001	7.75 0.0097
FSB		23.07 0.0001			38.20 0.0001	7.49 0.0108
FMB		23.56 0.0001			38.13 0.0001	7.85 0.0093

Table 3: Mean[a] of error rate and mean of *SSE*[b] for the main effects of 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
	0.356^b	0.422	0.507	0.271	0.420	0.425	0.322	0.550	0.228
M	0.247	0.254	0.269	0.232	0.322	0.231	0.126	0.250	0.251
	0.417	0.345	0.513	0.249	0.386	0.419	0.339	0.491	0.271
NS	0.183	0.176	0.175	0.184	0.260	0.184	0.095	0.179	0.180
	0.541	0.792	0.912	0.421	0.878	0.686	0.435	0.924	0.409
R	0.148	0.149	0.131	0.165	0.229	0.151	0.065	0.149	0.148
	0.952	1.182	1.544	0.590	1.386	1.110	0.704	1.384	0.749
U	0.248	0.251	0.266	0.232	0.359	0.252	0.136	0.248	0.250
	0.533	0.405	0.637	0.301	0.497	0.519	0.390	0.573	0.365
$\bar{U}$	0.273	0.262	0.292	0.243	0.382	0.271	0.151	0.267	0.268
	0.486	0.340	0.577	0.249	0.450	0.465	0.324	0.496	0.330
JK	0.241	0.248	0.259	0.230	0.354	0.247	0.132	0.244	0.245
	0.525	0.412	0.641	0.297	0.500	0.514	0.392	0.592	0.345
ISE	0.230	0.231	0.239	0.222	0.328	0.235	0.129	0.231	0.230
	0.377	0.393	0.505	0.265	0.408	0.425	0.322	0.535	0.235
IME	0.232	0.232	0.241	0.223	0.329	0.237	0.130	0.232	0.232
	0.378	0.388	0.501	0.265	0.409	0.419	0.322	0.532	0.234
ISC	0.227	0.229	0.236	0.220	0.325	0.233	0.127	0.228	0.228
	0.383	0.403	0.517	0.269	0.421	0.433	0.326	0.549	0.238
IMC	0.229	0.230	0.238	0.221	0.326	0.234	0.128	0.229	0.229
	0.384	0.397	0.512	0.269	0.421	0.426	0.324	0.545	0.236
ISB	0.232	0.233	0.242	0.224	0.330	0.237	0.131	0.233	0.233
	0.372	0.379	0.490	0.261	0.400	0.415	0.312	0.522	0.229
IMB	0.234	0.234	0.244	0.225	0.331	0.239	0.133	0.234	0.234
	0.369	0.373	0.482	0.259	0.393	0.410	0.310	0.512	0.229
FSE	0.232	0.232	0.242	0.223	0.330	0.237	0.130	0.232	0.232
	0.373	0.388	0.498	0.263	0.401	0.421	0.320	0.528	0.233
FME	0.233	0.233	0.243	0.223	0.330	0.238	0.131	0.233	0.233
	0.376	0.385	0.497	0.265	0.405	0.417	0.321	0.529	0.233
FSC	0.229	0.230	0.238	0.221	0.326	0.234	0.128	0.230	0.229
	0.379	0.398	0.509	0.268	0.413	0.428	0.324	0.541	0.236

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
	0.382	0.394	0.508	0.269	0.418	0.424	0.323	0.541	0.235
FSB	0.235	0.235	0.245	0.225	0.332	0.239	0.133	0.235	0.235
	0.369	0.373	0.482	0.259	0.392	0.410	0.310	0.514	0.227
FMB	0.235	0.235	0.245	0.225	0.332	0.240	0.133	0.235	0.235
	0.367	0.370	0.478	0.259	0.389	0.407	0.309	0.508	0.229

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. 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 experimental factors f and ν are highly important for all estimators with respect to the *SSE* 's; the effect due to the interaction p x ν is significant for all methods except *M* and $\bar{U}$; factor p is important only for *NS*, *R*, *U*, $\bar{U}$, and *JK* ; Δ^2 has significant effect only on *NS*, *R*, and $\bar{U}$; and only $\bar{U}$ is significantly affected by the interaction between p and f.

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 behaviour 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 *SSE* for the main effects of these four factors are presented in Table 3.

From Table 3, it is clear that the bootstrap 3.632 estimators (*ISE*, *IME* etc.) do not estimate the true error rate in the neighborhood of the interval (0.3, 0.4). It is also very prominent from this table that *R* and *NS* are the worst estimators not only with large *SSE* 's but also are 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 the balanced bootstrap estimators have smaller *SSE* 's than the other bootstrap methods for both cases of $p = 5$ and 10; the mixture sampling based estimators *FMB* and *IMB* being the best. However, *OS* estimator is better than the balanced bootstrap ones for $p = 5$, with the smallest *SSE* and becomes the best for this case. The behaviour of this estimator becomes worst when $p = 10$, and for this case the estimators that outperform the bootstrap estimators are $\bar{U}$ and *M*, $\bar{U}$ with the smallest *SSE*.

As far as the influence of the sample size factor f on the estimators is concerned, Table 3 shows that the average *SSE* from, *large samples* are much smaller than those from *small sample* cases for all estimators. For small f, *FMB* is the best (bootstrap and overall) estimator, while *IMB* and *FSB* are not far behind. For large samples, *M* and $\bar{U}$ are the overall best estimators with the smallest *SSE*, while *IMB*, *FSB*, and *FMB* perform better than the other bootstrap methods.

Now consider the behaviour of the estimation methods on the levels of the distance (separation) factor Δ^2. It is obvious from Table 3 that the *SSE* 's of all estimators for *highly separated* populations (with $\Delta^2 = 6.574$) are much smaller than those for $\Delta^2 = 2.836$ and $\Delta^2 = 1.098$. Although the *SSE* 's corresponding to $\Delta^2 = 2.836$ are consistently larger than those for $\Delta^2 = 1.098$, this difference is considerably small. This behaviour suggests that *SSE* does not decrease monotonically with increasing distance between populations. The balanced bootstrap estimators outperform all the other estimators in all cases except when $\Delta^2 = 1.098$ for which the *M* estimator has smallest *SSE*, though *IMB* and *FSB* are not so far behind. Among these balanced bootstrap estimators the *mixture* sampling versions (*IMB* and *FMB*) seem slightly better than the separate sampling ones (*ISB* and *FSB*).

Finally, from Table 3 we can easily deduce that the *SSE*'s for all methods are much smaller when the variables are almost independent ($\nu = 0.8$) than when the variables are highly interdependent ($\nu = 0.4$). This illustrates the high significance of the difference between the levels of factor ν in our *ANOVA*'s earlier. The estimator *M* closely followed by $\bar{U}$ are the best for the case $\nu = 0.4$, while *FMB* is the best choice among bootstrap estimators followed by *FSB* and *IMB*. Although *FSB* yields the smallest *SSE* for $\nu = 0.8$, the differences between the *SSE* 's for the estimators *OS*, *ISB*, *IMB*, and *FMB* are very small.

There are some interesting and peculiar behaviours to be noted from Table 3. It is clear that almost always the cross validation based estimators *U* and *JK* yield the largest *SSE* values, hence are the worst estimators in the sum-square-error sense for estimating the actual error rates. 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 *SSE* 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 *SSE* values, it was found that the influence of the interactions *METH* x p x f and *METH* x p x ν is much higher than those of the other interactions. Thus we may choose to interpret only the effects of p x f and p x ν on the 7 estimators considered. However, the individual *ANOVA*'s suggests that the significant influence of p x f on the *METH* factor may be due only to the $\bar{U}$ estimator. We also may argue that the interaction p x ν has significant influence only on 5 of the 7 estimators, namely, the *OS*, *U*, *FME*, *FMC* and *FMB* estimators. Hence, it would be appropriate to interpret the effect of p x f only on the $\bar{U}$ estimator, while the influence of p x ν should be examined only on the estimators *OS*, *U*, *FME*, *FMC* and *FMB*.

Result of the analysis shows that the $\bar{U}$ estimator has smaller *SSE* means when the sample sizes are large for both small and large number of variables in the data. This estimator also has smaller *SSE* means when $p = 10$ than when $p = 5$, for both levels of the sample sizes. These differences are larger for the cases with small samples than those with large samples. As far as the effect of interaction between the number of variables and interdependency of variables is concerned, result of the analysis shows that all the 5 estimators have smaller *SSE* means when the variables are independent than when they are interdependent, for both levels of p. These differences are greater when $p = 10$ than when $p = 5$. All the 5 estimators uniformly have the smallest *SSE* 's when the data consist of 10 independent variables, hence this combination becomes the best. We may conclude the *large number of independent variables* seem to reduce the *SSE* of the error rate estimators dramatically.

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 has the smallest *SSE* for independent variables. For all the other situations, the Finite Mixture Balanced (*FMB*) estimator is the best.

If we compare all estimators together, the *FMB* estimator is the best with smaller *SSE* values for small samples cases, or cases with medium and well separated populations. It is also becomes the overall best choice with minimum *SSE.* The overall situation refers to the behaviour averaged over all 48 Monte Carlo situations explained in section 3. The behaviour of the Infinite Mixture Balanced (*IMB*) and the Finite Separate Balanced (*FSB*) estimators are not far behind that of *FMB*. For the other situations, the best estimators are *OS* and $\bar{U}$ for small and large number of variables respectively, with the smallest *SSE*, while *M* and $\bar{U}$ are the best for large samples, *M* for close populations or interdependent variables, and *FSB* for independent variables.

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. (1974a). An Asymptotic Unbiased Techniques for Estimating The Error Rate in Discriminant Analysis, *Biometrics*, **30**, 239-249.

[14] McLachlan, G.J. (1974b). Estimation of The Error of Misclassification on the Criterion of Asymptotic Mean Square Error, *Technometrics*, **16**, 255-260.

[15] McLachlan, G.J. (1974c). The Relationship in Term of Asymptotic Mean Square Error Between The Separate Problems of Estimating Each of The Three Types of Error Rate of The Linear Discriminant Function, *Technometrics*, **16**, 569-574.

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

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

[18] Snapinn, S.M., and Knoke, J.D. (1984). Classification Error Rate Estimators Evaluated by Unconditional Mean Square Error, *Technometrics*, **26**, 371-378.

[19] Snapinn, S.M., and Knoke, J.D. (1985). An Evaluation of Smoothed Classification Error-Rate Estimators, *Technometrics*, **27**, 199-206.