Journal of the Korean
Statistical Society
Vol. 21, No. 1, 1992

A VARIABLE SELECTION IN HETEROSCEDASTIC DISCRIMINANT ANALYSIS : GENERAL PREDICTIVE DISCRIMINATION CASE[1]

Hea-Jung Kim[2]

ABSTRACT

This article deals with variable selection problem under a newly formed predictive heteroscedastic discriminant rule that accounts for multiple homogeneous covariance matrices across the K multivariate normal populations. A general version of predictive discriminant rule, a variable selection criterion, and a criterion for stopping with further selection are suggested. In a simulation study the practical utilities of those considered are demonstrated.

1. INTRODUCTION

Gnanadesikan and Kettenring(1984) classified discriminant analysis as one of the most useful statistical techniques for social problems. Widespread prevalence of the discriminant problem in many fields has seen the development of a plethora of new approaches for

[1]This paper was supported by NON DIRECTED RESEARCH FUND, Korea Research Foundation, 1990.

[2]Department of Statistics, Dongguk University, Seoul, 100-715, Korea

discriminant analysis. Among them, two major approaches, namely estimative and predictive methods, are well accepted and commonly used. Practical differences of them are illustrated by Aitchison and Donsmore(1975). Aitchison, et al.(1977) compared the two methods, and then advocated the use of predictive methods when the population distribution can be transformed to multinormality. They also suggested the use of heterosecdastic predictive method when there is a high possibility that the covariance matrices may differ appreciably across the populations.

In multiple discriminant analysis, it is common practice that once data set rejects the test for equality of covariance matrices we directly adopt the heteroscedastic discriminant analysis which assumes heterogeneous covariance matrices across the populations. In practice this is not always the case, since one would expect that measurement of the same characteristic in different populations would give rise to at least more than one very similar even if not identical covariance matrices groups, namely multiple homogeneities, among them(See Kim, 1991, for example). This gives rise to three major problems in discriminant analysis: problem of detecting the multiple homogeneities from a given data set, that of constructing discriminant rule for the case of the multiple homogeneities, and that of variables selection. Among them, the first problem can be circumvented by the test developed by Kim(1991). Thus, to make discriminant analysis possible even for the case of the multiple homogeneities, we need to solve the remaining two problems.

The purpose of this paper is to suggest a predictive method which not only resolves those remaining problems but also handles the homo and the heterogeneous covariances cases and to show the suggested method is likely to yield the more reliable results than the most commonly used method(heteroscedastic predictive method, Cf. Aitchison, et al., 1977). In the next section we develop a general version of predictive heteroscedastic discriminant rule that accounts for the multiple homogeneities. Under the possibility of the multiple homogeneities in the population distributions, some fundamental aspects of the discriminant variable selection problem such as the criterion for selection of variables and the criterion for stopping with further selection are considered and suggested in Section 3. In Section 4, by a simulation study, the suggested method is compared with the most commonly used method. The comparisons are done in terms of the performance of the discrimination and the variables selection under the multiple homogeneities. Some concluding remarks are given in Section 5.

2. A GENERAL PREDICTIVE DISCRIMINANT RULE

When we have a multiple homogeneities across the population covariance matrices, it is natural to use a discriminant rule which takes good care of the situation. Unfortunately, we don't have a formalized discriminant rule for the multiple homogeneities. In this

section we formulate a general predictive heteroscedastic discriminant rule, in a sense that, it is not only for homo and heteroscedastic discrimination but also for the multiple homogeneities.

Suppose we have K populations Π_i, $i = 1, \cdots, K$ each specified by a p-variate normal density $\eta(\cdot \mid \theta_i)$, $\theta_i = (\mu_i, \Sigma_i)$, with m multiple homogeneities across the covariance matrices so that underlying populations model is

$$M^* : \Sigma_1 = \cdots = \Sigma_{k_1}, \Sigma_{k_1+1} = \cdots = \Sigma_{k_1+k_2}, \cdots, \Sigma_{k^*} = \cdots = \Sigma_K,$$

$$k^* = \sum_{j=1}^{m-1} k_j, K = \sum_{j=1}^{m} k_j, \tag{2.1}$$

and suppose, for each of the populations, we have p $\times$ N_i training sample data matrix X_i based on N_i independent observations. If we let z be a new observation to be assigned to one of the K populations, with prior probability π_i of belonging to Π_i, $\sum_{i=1}^{K} \pi_i = 1$. Then the probability $\Pr(z \in \Pi_i \mid z, \Theta)$, that we would assign to population Π_i for a case with observation vector z, would be computed from Bayes' formula as

$$\Pr(z \in \Pi_i \mid z, \Theta) \propto \pi_i P(z \mid \Theta, \Pi_i), \text{ where } \Theta = \bigcup_{i=1}^{K} \theta_i. \tag{2.2}$$

In practice we never know Θ but we usually have training samples $\mathbf{X} = \bigcup_{i=1}^{K} X_i$. A method that replaces the probability density function $P(z \mid \Theta, \Pi_i)$ on the right side of (2.2) by

$$f(z \mid \mathbf{X}, \Pi_i) = \int_\Theta P(z \mid \Theta, \Pi_i) g(\Theta \mid \mathbf{X}) d\Theta, \tag{2.3}$$

has known to be as predictive method(Geisser, 1964; Fatti, et al., 1982, among others). Here $g(\Theta \mid \mathbf{X})$ denotes the posterior desity of Θ.

Lemma 1. If the underlying K populations model is M^*, and if the prior distribution of Θ is diffuse, the predictive density of z in (2.3) is

$$f(z \mid \mathbf{X}, \Pi_i) = St_p\{N(t) - k_t, \overline{X}_i, (1 + 1/N_i)S(k_t)\}, \tag{2.4}$$

$$a(t) \le i \le b(t), \; t = 1, \cdots, m,$$

where $\overline{X}_i$ and S_i are sample mean and covariance matrices of i-th training sample, respec-

tively and $N(t) = \sum_{i=a(t)}^{b(t)} N_i, S(k_t) = \sum_{i=a(t)}^{b(t)} (N_i - 1)S_i/(N(t) - k_t), a(t) = \sum_{j=1}^{t-1} k_j + 1, b(t) = \sum_{j=1}^{t} k_j$. Here $St_p(\cdot,\cdot,\cdot)$ denotes a p-dimensional variate t density function defined on X = R^p by the density at z (Cf. Kim, 1991).

Proof. If we let $\Sigma_{a(t)} = , \cdots , = \Sigma_{b(t)} = \Sigma(k_t)$ $t = 1, \cdots, m$, under the diffuse prior

$$P_0(\Theta) \propto \prod_{t=1}^{m} \mid \Sigma(k_t) \mid^{-(p+1)/2}, \tag{2.5}$$

the posterior density of Θ is

$$g(\Theta \mid \mathbf{X}) \propto \prod_{t=1}^{m} \mid \Sigma(k_t) \mid^{-(N(t)-p-1)/2} \; exp\{-\frac{1}{2}\, tr\{\Sigma(k_t)^{-1}C(k_t)\}\}, \tag{2.6}$$

where $C(k_t) = (N(t) - k_t)S(k_t) + \sum_{j=a(t)}^{b(t)} N_j(\bar{X}_j - \mu_j)(\bar{X}_j - \mu_j)'$.

Without loss of generality, we assume that the probability density $P(z \mid \Theta, \Pi_i) = \eta(z \mid \mu_i, \Sigma(k_1))$ such that $\Sigma_i = \Sigma(k_1)$. Then, from the definition (2.3), integrating out all parameters Θ except for $\theta_i = (\mu_i, \Sigma(k_1))$ that relate to the distribution of z, we have

$$f(z \mid \mathbf{X}, \Pi_i) \propto \int\int \mid \Sigma(k_1) \mid^{-(N(1)-k_1+p+3)/2} exp\{-\tfrac{1}{2} tr[\Sigma(k_1)^{-1}\mathbf{A_i}]\partial\Sigma(k_1)\partial\mu_i,$$

where $\mathbf{A_i} = (N(1) - k_1)S(k_1) + N_i(\bar{X}_i - \mu_i)(\bar{X}_i - \mu_i)' + (z - \mu_i)(z - \mu_i)'$.

Noting that if the integrand is viewed as a function of $\Sigma(k_1)$, it is proportional to a inverted Wishart density(Cf. Press, 1982), the integration over $\Sigma(k_1)$ is

$$f(z \mid \mathbf{X}, \Pi_i) \propto \mid \mathbf{B_i} \mid^{-(N(1)-k_1+2)/2} \int \{1 + (\mu_i - \bar{\mu}_i)'\mathbf{B_i}^{-1}(\mu - \bar{\mu}_i)\}^{-(N(1)-k_1+2)/2}\partial\mu_i,$$

where $(N_i + 1)\mathbf{B_i} = (N(1) - k_1)S(k_1) + N_i(z - \bar{X}_i)(z - \bar{X}_i)'/(N_i + 1)$.
Now the integrand is proportional to a p-dimensional variate t density. Integrating gives

$$f(z \mid \mathbf{X}, \Pi_i) \propto \{1 + \tfrac{N_i}{(N_i+1)(N(1)-k_1)}(z - \bar{X}_i)'S(k_1)^{-1}(z - \bar{X}_i)\}^{-(N(1)-k_1+1)/2}. \tag{2.7}$$

This is the Kernel of $St_p\{N(1) - k_1, \bar{X}_i, (1 + 1/N_i)S(k_1)\}$, so that if we apply this result to the general z from Π_i with covariance $\Sigma_i = \Sigma(k_t)$, $a(t) \leq i \leq b(t)$, $t = 1, \cdots, m$, we get (2.4). Therefore we prove the lemma.♠

Corollary 1. Under the assumptions of lemma 1, if a prior probability of belonging to i-th population Π_i is π_i, $\sum_{i=1}^{K} \pi_i = 1$, the predictive discriminant rule is to classify a new observation z into a population which satisfies

$$\mathcal{M}ax \;\; \pi_i f(z \mid \mathbf{X}, \Pi_i), \quad i = 1, \cdots, K. \tag{2.8}$$

Proof. Apply Bayes theorem to the predictive probability density (2.4) for classifying a new observation z into Π_i. This gives the posterior probability of z belonging to Π_i;

$$Pr(z \in \Pi_i \mid z, \pi_i) \;\propto\; \pi_i f(z \mid \mathbf{X}, \Pi_i). \tag{2.9}$$

Thus (2.8) is equivalent to the usual posterior odds ratio criterion for optimum classification(Cf. Geisser, 1964).♠

The results above can be applicable for any type of multiple homogeneites among K population covariance matrices if we reconstruct(or exchange) the subscripts involved in the multiple homogeneities at hand to conform with those of M^*. Furthermore, it can be easily shown that (2.4) and (2.8) give exactly the same homo and heteroscedastic predictive discriminant rule(Cf. Aitchison, et al. , 1977) when we put M^* with sets of subscripts $k_1 = \mathrm{K}$, $k_2 = \cdots = k_m = 0$ and $k_1 = \cdots = k_m = 1$, respectively. Particularly, in the latter case, (2.4) gives the predictive density used for heteroscedastic predictive discriminant rule;

$$f(z \mid \mathbf{X}, \Pi_i) = St_p\{N_i - K, \overline{X}_i, (1 + 1/N_i)S_i\}. \tag{2.10}$$

3. VARIABLE SELECTION

In any application of discriminant analysis, some variables will show greater variation between populations, relative to their variation within populations. The question then arises how we can determine the individual contribution of variables to the over all discrimination so that we can drop any of those variables from the analysis without appreciably increasing the error rate. Since Rao(1965) provided the largest F criterion for additional discrimination, there have been many variable selection criteria for multiple discriminant analysis. Among them, Wilk's Λ, minimum Mahalanobis' D^2, likelihood ratio(Cf. Fatti, et al. 1982), and error rate criterion (Cf. Habbema and Hermans, 1977) can be illustrated. However, those criteria are based upon the assumptions that covariance

matrices are homogeneous or heterogeneous across the populations. Accordingly, except for an error rate criterion considered below, those are not applicable to our variable selection situation that assumes the multiple homogeneities among the population covariance matrices.

If we let $Pr(\Pi_i \mid \Pi_i)$ be the predictive probability that a new observation z has been classified as belonging to Π_i when in fact it belongs to Π_i, under the classification rule in (2.8), we have a predictive probability of misclassification(the error rate) defined by

$$E = 1 - \sum_{i=1}^{K} Pr(\Pi_i \mid \Pi_i), \tag{3.1}$$

where

$$Pr(\Pi_i \mid \Pi_i) = \pi_i \int_{R_i} f(z \mid \mathbf{X}, \Pi_i) dz,$$

$$R_i = \{z : \pi_i f(z \mid \mathbf{X}, \Pi_i) > \pi_\ell f(z \mid \mathbf{X}, \Pi_\ell),\ i \neq \ell\},$$

the classification region to Π_i, $i,\ \ell = 1, \cdots, K$.

In order to find the optimal subset of variables when using the error rate criterion(3.1), one has to compute the estimated rate of correct classification $\hat{P}_c = \sum_{i=1}^{K} \hat{P}r(\Pi_i \mid \Pi_i)$ for all subsets of p discriminant variables, and choose the subset having maximal estimated rate of correct classification. This is called exhaustive enumeration procedure for the variable selection. When the investigation of all possible subsets takes too much computer time, a stepwise approach can be used leading to a suboptimal subset.

To estimate the error rate E on the basis of the training samples $\mathbf{X} = \bigcup_{i=1}^{K} X_i$, the most obvious estimate is the apparent error rate which is the proportion of observed errors made by (2.8) on its own training samples $\mathbf{X}$. Usually the apparent error rate tends to be smaller than the error rate(Cf. Efron, 1983), because the same data have been used both to construct and evaluate (2.8). The leaving-one-out method(a cross-validation) circumvents this difficulty by removing each observation from the data set used in its own discrimiation. Let $X(ij)$ be the training samples with j-th observation vector x_{ij} of i-th data matrix X_i removed, and let Y_{ij} indicate the correctness of classifying x_{ij} based upon the discriminant rule (2.8);

$$Y_{ij} = \begin{cases} 0 & \text{if } max\ \pi_\ell f(x_{ij} \mid X(ij), \Pi_\ell) = \pi_i f(x_{ij} \mid X(ij), \Pi_i) \\ 1 & \text{otherwise.} \end{cases}$$

Then the estimated error rate based on the leaving-one-out method is

$$\hat{E} = \frac{1}{N} \sum_{i=1}^{K} \sum_{j=1}^{N_i} Y_{ij}, \tag{3.2}$$

where $\hat{E} = 1 - \hat{P}_c$, $i = 1, \cdots, K$, $j = 1, \cdots, N_i$, $N = \sum_{i=1}^{K} N_i$. A good reference of this method is Lachenbruch and Mickey(1968), and Efron(1983) showed that it provides a nearly unbiased estimate of the error rate which is closely related to bootstrap estimate of the error rate. An advantage of the error rate criterion, besides its direct interpretation, is that it naturally provides a stopping criterion. Formally, this criterion stops selection at step q if, for a given threshold value Δ,

$$\hat{P}_c(S_j) - \hat{P}_c(S_{j-1}) \geq \Delta, \quad j = 1, \cdots, q \tag{3.3}$$

while

$$\hat{P}_c(S_{q+1}) - \hat{P}_c(S_q) < \Delta,$$

where the subset S_j is optimal with respect to the rate of correct classification among all subsets of size j considered, and $\hat{P}_c(S_j)$ denotes the corresponding estimated correct classification rate.

4. SIMULATION STUDIES

A sampling experiments was performed to compare the performance of the two heteroscedastic predictive methods under the presence of the multiple homogeneities: one in the most commonly used method(heteroscedastic predictive method by Aitchison, et al., 1977) which estimates the probability density function in (2.3) by (2.10) and the other by (2.4) as suggested in Section 2. We will call the latter method as the suggested predictive method.

For a comparison we consider 21 discrimination situations with three populations $\Pi_i \sim N_p(\mu_i, \Sigma_i)$, i = 1, 2, 3, with $\Sigma_1 \neq \Sigma_2 = \Sigma_3 = \Sigma$. Using a non-singular linear

transformation H such that $H\Sigma_1 H' = I_p$ and $H\Sigma H = D$, a diagonal matrix, 30 samples for each of simulation characterized by the set

$$\{\mu_1, \mu_2, \mu_3, H, D, p, N_1 = N_2 = N_3 = J\}$$

were generated. These are done with three different experiments designed to see the discrimination effect of location parameters. These experiments are set by changing the values of population mean vectors;

Table 1. Population Mean Vectors

Mean Vectors	μ_1	μ_2	μ_3
Experiment 1	(0, -3, 0, -3, 0, -3, 0)'	(-3, 0, -3, 0, -3, 0, -3)'	(1, 1, 1, 1, 1, 1, 1)'
Experiment 2	(0, -3, 0, -3, 0, -3, 0)'	(-3, 5, -3, 5, -3, 5, -3)'	(1, 1, 1, 1, 1, 1, 1)'
Experiment 3	(0, -3, 0, -3, 0, -3, 0)'	(-3, 5, -3, 5, -3, 5, -3)'	10 x (1, 1, 1, 1, 1, 1, 1)'

Table 2 provides details of the simulation situations in each experiment which has been used to assess the relative merits of the two methods.

Table 2. Simulated Situations Used in Each Experimental Comparison

p	J	H	D	μ_1	μ_2	μ_3
3	5, 10, 20	Q_3	D_3	$\mu_{1(3)}$	$\mu_{2(3)}$	$\mu_{3(3)}$
5	10, 20	Q_5	D_5	$\mu_{1(5)}$	$\mu_{2(5)}$	$\mu_{3(5)}$
7	10, 20	Q	D	μ_1	μ_2	μ_3

Note: D = diag(2, 3, 4, 5, 6, 7, 8), Q is the same matrix used in Press(1982, p.275).
Note: D_i and Q_i, i = 3, 5, denote principal submatrices with first i diagonal elements of D and H respectively and $\mu_{1(i)}$, $\mu_{2(i)}$, and $\mu_{3(i)}$ are i x 1 subvectors composed of first i components of μ_1, μ_2, and μ_3, respectively.

4.1 Comparison of the Two Methods

Each simulated training set, say $\{x_{hij} : h = 1, \cdots, 30, i = 1, \cdots, 3, j = 1, \cdots, J\}$ gives rise to two assessments of log differences: the differences between the true one p(z | Θ, Π_i) in (2.2) and one for each of the two in (2.4) and (2.10). The measure of performance of each predictive function is the mean log absolute deviation(MLAD) from the true one

$$MLAD = \frac{1}{90J} \sum_{h,i,j} log \mid f(x_{hij} \mid \mathbf{X}, \Pi_i) - p(x_{hij} \mid \Theta, \Pi_i) \mid, \tag{4.1}$$

Since the measure (4.1) is invariant with respect to location parameter, we calculate this measure based only on the first experiment, and results are given in Table 3.

Table 3. The Performance of The Two Predictive Functions

p	J	MLAD1	MLAD2	MLAD1-MLAD2
3	5	-6.979	-6.901	-.078
	10	-7.130	-7.097	-.033
	20	-6.692	-6.668	-.024
5	10	-11.571	-11.446	-.125
	20	-10.721	-10.679	-.042
7	10	-16.264	-16.110	-.154
	20	-15.111	-15.023	-.088

Note: MLAD1 denotes MLAD calculated by the general predictive density in (2.4).
Note: MLAD2 denotes MLAD calculated by the heteroscedastic predictive density in (2.10).

Table 3 shows that the values of MLAD1 are uniformly smaller than those of MLAD2. This implies that, under a multiple homogeneities, the general predictive density (2.4) better estimates the true one in (2.2). The two discriminant methods are then compared in terms of the error rates(misclassification rates). The error rates for each method are estimated by the leaving-one-out method. The advantage of this estimation method is well illustrated in Lachenbruch and Micky(1968).

Table 4. Estimated Error Rates (averaged over 30 replicates)

p	J	Experiment 1		Experiment 2		Experiment 3	
		method 1	method 2	method 1	method 2	method 1	method 2
	5	.505	.731	.345	.627	.120	.495
3	10	.395	.464	.271	.321	.063	.147
	20	.308	.311	.188	.213	.040	.085
5	10	.348	.589	.130	.286	.008	.157
	20	.230	.331	.052	.111	.005	.060
7	10	.243	.586	.045	.205	.002	.169
	20	.224	.340	.026	.026	.005	.045

Note: Method 1 denotes the suggested method using (2 . 4) and method 2 denotes the heteroscedastic predictive method using (2 . 10). Note: Estimated error rate = (number of misclassified samples by leaving-one- out method) / (total number of generated samples in each experiment situation).

Several points were noted in constructing this table. First, for all simulation situations, the estimated error rate of the suggested predictive method is lower than that of the most commonly used method. Second, differences in the error rates between them get bigger when sample size gets smaller. Finally, the suggested method gives better results when dimensionality is large and did better when centroid distance between populations gets bigger.

4.2 Comparison in Variables Selection

When multiple homogeneities are presented in the covariance matrices, as demonstrated by the preceding examples, the suggested predictive method(method 1) performs better than the most commonly used method(method 2). Now we further develop a study to see difference in the variable selection between the two methods by estimating the error rate in (3.1).

The following example is based on a sampling experiment generated from the simulation situation with p=7 and J=20 in the first experiment. So that these data consist of

three populations with 20 sample elements per population and 7 variables. For each of the two predictive methods, using the stepwise procedure and the total enumeration procedure, the subsets of variables are selected at each stage. The criterion (3.3) for stopping with further selection of variables is used for the stepwise procedure. Table 5 and Table 6 give "stepwise results" and "exhaustive enumeration results", respectively. Here we used the leaving-one-out method to calculate $\hat{P}_c(S_j)$'s for (3.3) and the stepwise procedure was forced to take seven selection steps by putting Δ = -1.

Table 5. Variables Selected by Stepwise Procedure

	Method 1		Method 2	
Stage(j)	Variables	$\hat{E}(S_j)$	Variables	$\hat{E}(S_j)$
1	6	.367	7	.367
2	6 7	.300	6 7	.283
3	3 6 7	.250	1 6 7	.267
4	1 3 6 7	.250	1 4 6 7	.250
5	1 3 4 6 7	.183	1 4 5 6 7	.250
6	1 2 3 4 6 7	.200	1 2 4 5 6 7	.333
7	1 2 3 4 5 6 7	.200	1 2 3 4 5 6 7	.333

Note: estimated error rate $\hat{E}(S_j) = 1 - \hat{P}_c(S_j), j = 1, \cdots, 7$.

Table 6. Variables Selected by Exhaustive Enumeration Procedure

	Method 1		Method 2	
Subset Size(j)	Variables	$\hat{E}(S_j)$	Variables	$\hat{E}(S_j)$
1	6	.367	7	.367
2	2 4	.283	6 7	.283
3	1 2 4*	.167	1 2 7*	.217
4	1 2 3 5	.183	1 4 6 7	.250
5	1 2 3 4 7	.167	1 4 5 6 7	.250
6	1 2 3 4 5 7	.183	1 2 3 4 6 7	.250
7	1 2 3 4 5 6 7	.200	1 2 3 4 5 6 7	.333

Note: * denotes the optimal subset of each method.

This example shows that the most commonly used method(heteroscedastic predictive method) misleads the variable selection when the multiple homogeneities exist across the

populations. Also, the estimated error rates in the tables evidently reveal the superior performance of the suggested predictive method in selecting variables.

5. CONCLUDING REMARKS

When a discriminant situation requires the statistician to provide plausibility assessment for the possible populations, we have seen that total ignorance of the multiple homogeneities across the covariance matrices may lead to severe distortion of the assessment. Noticing that the suggested predictive method also embraces homo and heteroscedastic predictive discriminant procedures, the theoretical consideration of Section 2 and 3 ,and limited but comparative simulation studies of Section 4 can be considered as strong supports for the use of the suggested method.

However the suggested method is only applicable to the case when the exact structure of the multiple homogeneities among K covariance matrices is known beforehand so that we can construct M^* in (2.1). This problem may be easily solved if we adopt the multiple homogeneities test by Kim(1991) in the first stage of discriminant analysis.

ACKNOWLEDGEMENT

The detailed and constructive comments of referees were very useful in making the paper more readable. Their efforts are greatly appreciated.

REFERENCES

(1) Aitchison, J. and Dunsmore, I. R. (1975). *Statistical Prediction Analysis*, Cambridge University Press.

(2) Aitchison, J., Habbema, J. D. F. and Kay, J. W. (1977). A critical comparison of two methods of statistical discrimination. *Applied Statistics*, 26, No. 1, 15-25.

(3) Bartlett, M. S. (1937). *Properties of Sufficiency and Statistical Tests*, Proceedings of the Royal Society of London(A) 160, 268-282.

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

(5) Fatti, L. P., Hawkins, D. M. and Raath, E. L. (1982). *Topics in Applied Multivariate Analysis*, Ed by D. M. Hawkins, 1-71, Cambridge University Press.

(6) Geisser, S. (1964). Posterior odds for multivariate normal classifications. *Journal of Royal Statistical Society.* B, 26, 69-76.

(7) Gnanadesikan, R. and Kettenring, J. R. (1984). *A Pragmatic Review of Multivariate Methods in Applications.* In Statistics : An appraisal (H. A. David and H. T. David, eds.), 309-337. Iowa State University Press.

(8) Habbema, J. D. F. and Hermans, J. (1977). Selection of variables in discriminant analysis by F-statistic and error rate. *Technometrics*, Vol. 19, 487-493.

(9) Kim, H. J. (1991). A test for equality form of covariance matrices of multivariate normal populations. *Journal of Korean Statistical Society*, Vol. 2.

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

(11) Press, S. J. (1982). *Applied Multivariate Analysis*, Krieger Publishing Company, Florida.

(12) Rao, C. R. (1965). *Linear Statistical Inference and its Applications*, Wiley, New York.