Journal of the Korean Statistical Society (2000), 29: 3, pp 337–350

A Bayesian Diagnostic Measure and Stopping Rule for Detecting Influential Observations in Discriminant Analysis

Myung-Cheol Kim[1] and Hea-Jung Kim[2]

ABSTRACT

This paper suggests a new diagnostic measure and a stopping rulefor detecting influential observations in multiple discriminant analysis (MDA). It is developed from a Bayesian point of view using a default Bayes factor obtained from the fractional Bayes factor methodology. The Bayes factor is taken as a discriminatory information in MDA. It is shown that the effect of an observation over the discriminatory information is fully explained by the diagnostic measure. Based on the measure, we suggest a stopping rule for detecting influential observations in a given training sample. As a tool for interpreting the measure a graphical method is used. Performance of the method is examined through two illustrative examples.

Keywords: Multiple discriminant analysis; Influential observations; Discriminatory information; Fractional Bayes factor; Diagnostic measure; Stopping rule

1. Introduction

In practical applications of MDA with K multivariate normal populations, it is seldom wise to compute and report only the linear discriminant function. Generally one would wish to guard against, and check for, the possibility that some observations do not contributes to discrimination of the populations. For this purpose, many articles have been suggested diagnostic measures for the identification of outliers and influential observations in discriminant analysis. Critchley and Vitiello (1991) and Fung (1992) independently proposed two fundamental statistics, in Fisher's linear discriminant analysis (LDA), like the residual and leverage measure in regression, on which many influence measures depend. By means of the fundamental statistics, Critchley and Vitiello (1991) examined the

[1]Department of Industrial Engineering, Samchok National University, Kwangwon-do, Korea. 245-711

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

influence of observations upon misclassification probability estimates in LDA and Fung (1995) suggested a couple of Cook's type diagnostic measures for detecting outliers. For further references, see Johnson (1987) and Rencher (1995) and references therein.

The studies mentioned above are mainly designed for two-group discriminant analysis. However, as pointed out in Fung (1999), a method for detecting influential observations in MDA has not been seen yet. The present paper considers, however, a diagnostic measure for the influence of observations that can be applicable to both two-group discriminant analysis and MDA. The diagnostic measure is designed to detect observations influential on discriminatory information. It is developed by use of a default Bayes factor and related to the conditional predictive ordinate (cf. Pettit and Young 1990). Section 2 derives the Bayes factor via a development of the fractional Bayes factor method introduced by O'Hagan (1995) and justifies the use of the Bayes factor as a measure of discriminatory information in MDA. Based upon the information, Section 3 proposes a diagnostic measure for detecting influential observations in a training sample and develops a stopping rule for the detection. In Section 4 the performance of the proposed measure and the stopping rule is examined through two illustrative examples. It also proposes a way of visual interpretation, especially by use of a graphical method. A few concluding remarks are give in Section 5.

2. Discriminatory Information

Suppose we have K multivariate normal populations, $\Pi_1, \ldots, \Pi_K$ each specified by a model M_i, $i = 0, 1$, where M_i defines the distribution of each population distribution $\Pi_k \sim N_p(\mu_k, \Sigma)$, $k = 1, \ldots, K$. Let our interest of model comparison be homogeneity (or heterogeneity) of the mean vectors among K populations, and let the model specification be $\mu_1 = \cdots = \mu_K = \mu$ under M_0 and under M_1, $\mu_1 \neq \cdots \neq \mu_K$.

We suppose now that $X_1(k), \ldots, X_{N_k}(k)$ denote independent p variate sample of size N_k from Π_k with distribution $N_p(\mu_k, \Sigma)$, $k = 1, \ldots, K$, and suppose all the independent samples (so called the training sample) as D. Then the training sample D is to have arisen under one of the two models according to respective probability densities given by

$$f(D|\mu, \Sigma, M_0) = (2\pi)^{-\frac{Np}{2}} |\Sigma|^{-\frac{N}{2}} \exp\left\{ -\frac{1}{2} \mathrm{tr}[\Sigma^{-1}\Omega] \right\}, \qquad (1)$$

$$f(D|\mu_1,\cdots,\mu_K,\Sigma,M_1) \;=\; (2\pi)^{-\frac{Np}{2}}|\Sigma|^{-\frac{N}{2}}\exp\Big\{-\frac{1}{2}\mathrm{tr}\Big[\Sigma^{-1}\textstyle\sum_{k=1}^{K}\Omega_k\Big]\Big\}, \quad (2)$$

where $\Omega = V + N(\mu - \bar{X})(\mu - \bar{X})'$, $\Omega_k = V_k + N_k(\mu_k - \bar{X}(k))(\mu_k - \bar{X}(k))'$, $\bar{X}(k) = \sum_{j=1}^{N_k} X_j(k)/N_k$, $\bar{X} = \sum_{k=1}^{K} N_k\bar{X}(k)/N$, $N = \sum_{k=1}^{K} N_k$, $V_k = \sum_{j=1}^{N_k}(X_j(k) - \bar{X}(k))(X_j(k) - \bar{X}(k))'$, and $V = \sum_{k=1}^{K}\sum_{j=1}^{N_k}(X_j(k) - \bar{X})(X_j(k) - \bar{X})'$.

To derive a measure of group separation (or equivalently discriminatory information), our interest focuses primarily on a statement concerning to relative probability that D comes from one or the other of the model, and not about making probability statement about where a parameter lies. Therefore, we shall use a particular convenient prior densities, π_0 and π_1, to reflect an initial diffuseness or vagueness about the unknown parameters (cf. Jeffreys 1961);

$$\pi_0(\mu,\Sigma|M_0) \propto |\Sigma|^{-\frac{p+1}{2}}, \quad \pi_1(\mu_1,\cdots,\mu_K,\Sigma|M_1) \propto |\Sigma|^{-\frac{p+1}{2}}. \quad (3)$$

Using the definition of the marginal likelihood (cf. Kass and Raftery 1995), we may obtain, under the vague priors, respective marginal likelihoods conditionally on M_0 and M_1 that include undefined constants. Thus the use of the improper priors (3) leads to well known problem called arbitrariness of Bayes factor (cf. Berger and Pericchi, 1993). Various approaches have been advocated for dealing with this problem. One is to remove the indeterminacy by a kind of thought experiment as proposed by Spiegelhalter and Smith (1982). Another approach to improper priors makes use of a training sample. This includes the partial Bayes factors by Lempers (1971), the fractional Bayes factor by O'Hagan (1995), and the intrinsic Bayes factor of Berger and Pericchi (1996). Among them we will follow O'Hagan (1995) in the use of a proportion b of the data to resolve the problem of arbitrariness in the Bayes factor. The reason for adopting the fractional Bayes factor is due to its good properties such as consistency, simplicity, robustness and coherence (cf. O'Hagan, 1995).

Definition 1 (O'Hagan 1995). Let $f(D|\theta_i, M_i)$ be the full likelihood based upon model M_i with parameter vector θ_i, and if the prior density of θ_i has an improper form denoted by $\pi(\theta_i|M_i)$. Then, for $b \in (0,1)$,

$$B_{ij}^{b} = P_b(D|M_i)/P_b(D|M_j), \quad i \neq j$$

is referred to as an fractional Bayes factor of M_i relative to M_j, where

$$P_b(D|M_i) = \frac{\int \pi(\theta_i|M_i) f(D|\theta_i, M_i) d\theta_i}{\int \pi(\theta_i|M_i) f(D|\theta_i, M_i)^b d\theta_i}.$$

Using the definition we can eliminate the indeterminacy of the Bayes factor of M_0 relative to M_1 (cf. O'Hagan 1995).

Lemma 1. Under the likelihoods (1) and (2), and the improper priors (3) for the parameters in M_0 and M_1, the fractional Bayes factor method yields respective marginal likelihoods

$$\begin{aligned} P_b(D|M_0) &= b^{pNb/2}\pi^{(b-1)pN/2}|V|^{N(b-1)/2}\frac{\Gamma_p\{(N-1)/2)\}}{\Gamma_p\{(bN-1)/2)\}}, \\ P_b(D|M_1) &= b^{pNb/2}\pi^{(b-1)pN/2}|\sum_{k=1}^{K} V_k|^{N(b-1)/2}\frac{\Gamma_p\{(N-K)/2)\}}{\Gamma_p\{(bN-K)/2)\}}, \end{aligned}$$

where $b \in (0,1)$ and $\Gamma_p\{\theta\} = \pi^{p(p-1)/4}\prod_{j=1}^{p}\Gamma\{\theta-(j-1)/2\}$, a p-variate gamma function.

Proof. Using the fractional Bayes factor in Definition 1, we have the marginal likelihoods under M_0 and M_1 as

$$P_b(D|M_0) = \frac{\int(2\pi)^{-Np/2}|\Sigma|^{-(N+p+1)/2}\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\Omega]\right\}d\mu d\Sigma}{\int(2\pi)^{-Nbp/2}|\Sigma|^{-(Nb+p+1)/2}\exp\left\{-\frac{b}{2}tr[\Sigma^{-1}\Omega]\right\}d\mu d\Sigma}$$

and

$$P_b(D|M_1) = \frac{\int(2\pi)^{-Np/2}|\Sigma|^{-(N+p+1)/2}\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\sum_{k=1}^{K}\Omega_k]\right\}\prod_{k=1}^{K}d\mu_k d\Sigma}{\int(2\pi)^{-Nbp/2}|\Sigma|^{-(Nb-K+p+1)/2}\exp\left\{-\frac{b}{2}tr[\Sigma^{-1}\sum_{k=1}^{K}\Omega_k]\right\}\prod_{k=1}^{K}d\mu_k d\Sigma},$$

respectively. Noting that if the integrands of $P_b(D|M_1)$ are viewed as functions of μ_k they are proportional to p-variate normal densities. Completing the square in μ_k and integrations over μ_k give

$$P_b(D|M_1) = b^{pK/2}(2\pi)^{pN(b-1)/2}\frac{\int|\Sigma|^{-(N-K+p+1)/2}\exp\left\{-\frac{1}{2}tr[\Sigma^{-1}\sum_{k=1}^{K}V_k]\right\}d\Sigma}{\int|\Sigma|^{-(Nb+p+1)/2}\exp\left\{-\frac{b}{2}tr[\Sigma^{-1}\sum_{k=1}^{K}V_k]\right\}d\Sigma}.$$

Now the integrands are a function of Σ and they are proportional to inverted Wishart densities, the integrations over Σ on both numerator and denominator are easily found to result in $P_b(D|M_1)$. Similar proof holds for the derivation of $P_b(D|M_0)$.

Theorem 1. The fractional Bayes factor, B_{01}^{b}, of M_0 relative to M_1 is given by

$$B_{01}^{b} = \frac{\Gamma_p\{(bN-K)/2)\}}{\Gamma_p\{(N-K)/2)\}}\frac{\Gamma_p\{(N-1)/2)\}}{\Gamma_p\{(bN-1)/2)\}}\Lambda^{-\frac{N(b-1)}{2}}, \tag{4}$$

where $b \in (0,1)$ and $\Lambda = \left|\sum_{k=1}^{K} V_k\right| / |V| \sim \Lambda_{p,\alpha,\beta}$, a Λ distribution with $\alpha = K-1, \beta = N-K$ under M_0 (cf. Anderson 1984).

Proof. Since $B_{01}^b = P_b(D|M_0)/P_b(D|M_1)$, substituting the result of Lemma 1 gives the Bayes factor.

It is noted that the Bayes factor (4) is closely related to MANOVA test statistic, Wilks' Λ used for finding linear combinations of variables, i.e. linear discriminant function(LDF), that best separates groups of multivariate normal observations (cf. Rencher 1995, p 311). Therefore, we can take B_{01}^b as an information that measures a degree of group separation among multivariate normal observations used in MDA and call it as "discriminatory information" in MDA. Following remarks show that B_{01}^b can be taken as the discriminatory information.

Remark 1. In case $p = 1$, B_{01}^b is a function of F test statistic of one-way ANOVA so that

$$B_{01}^b = \frac{\Gamma\{(bN-K)/2)\}}{\Gamma\{(N-K)/2)\}} \frac{\Gamma\{(N-1)/2)\}}{\Gamma\{(bN-1)/2)\}} \left(1 + \frac{K-1}{N-K} F\right)^{\frac{N(b-1)}{2}},$$

where $F = (N-K)\sum_{k=1}^{K} N_k(\bar{X}(k) - \bar{X})^2 / ((K-1)\sum_{k=1}^{K}\sum_{i=1}^{N_k}(X_i(k) - \bar{X}(k))^2)$ that follows F distribution with degrees of freedom $K-1$ and $N-K$.

Proof. Substituting $p = 1$ in the expression of B_{01}^b in Theorem 1, we have the result.

Remark 2. In case $K = 2$ (i.e. Two-group discriminant analysis case), the Bayes factor reduces to a function of Hotelling's T^2 statistic:

$$B_{01}^b = \frac{\Gamma_p\{(bN-2)/2)\}}{\Gamma_p\{(N-2)/2)\}} \frac{\Gamma_p\{(N-1)/2)\}}{\Gamma_p\{(bN-1)/2)\}} \left(1 + \frac{T^2}{N-2}\right)^{\frac{N(b-1)}{2}},$$

where $T^2 = (N_1N_2/N)(\bar{X}(1) - \bar{X}(2))' S^{-1} (\bar{X}(1) - \bar{X}(2))$, $S = (V_1 + V_2)/(N-2)$, is Hotelling's T^2 statistic for comparing M_0 with M_1.

Proof. For $K = 2$, the Bayes factor in Theorem 1 reduces to $B_{01}^b = \Gamma_p\{(bN-2)/2)\}\Gamma_p\{(N-1)/2)\}(|V|/|V_1+V_2|)^{-N(b-1)/2}/(\Gamma_p\{(N-2)/2)\}\Gamma_p\{(bN-1)/2)\})$. Since $V = V_1 + V_2 + \sum_{k=1}^{2} N_k(\bar{X}(k) - \bar{X})(\bar{X}(k) - \bar{X})'$, we have

$$\begin{aligned} |V|/|V_1+V_2| &= |V(V_1+V_2)^{-1}| \\ &= \{1 + (N_1N_2/N)(\bar{X}(1) - \bar{X}(2))'(V_1+V_2)^{-1}(\bar{X}(1) - \bar{X}(2))\}. \end{aligned}$$

Applying the definition of Hotelling's T^2-statistic in the last term and expressing B_{01}^b in terms of the T^2, we have the result.

Remark 2 is of particular interest as providing a Bayesian alternative to the two-sample Hotelling's T^2-test. Furthermore, when $p = 1$ and $K = 2$, B_{01}^b becomes a function of two sample t-test statistic for testing M_0 against M_1 such that

$$B_{01}^b = \frac{\Gamma\{(bN-2)/2)\}}{\Gamma\{(N-2)/2)\}} \frac{\Gamma\{(N-1)/2)\}}{\Gamma\{(bN-1)/2)\}} \left(1 + \frac{t^2}{N-2}\right)^{\frac{N(b-1)}{2}},$$

where $N = N_1 + N_2$ and t follows t distribution with $N-2$ degrees of freedom.

Remark 3. In discriminant analysis, as the training sample D contains the larger information of group separation, i. e. discriminatory information in MDA, the value of B_{01}^b becomes the smaller.

Proof. In the foregoing results we have seen that the Bayes factor B_{01}^b is directly related with classical measures of discriminatory information. It has inverse relation with Hotelling's T^2 for two-group LDA, while it is proportional to Wilks' Λ statistics for MDA (cf. Hawkins, 1982 and McLachlan, 1992), and hence these relations give the result.

The key question remaining in the use of B_{01}^b as a measure of discriminatory information in D is the choice of proper b. For the choice of b, O'Hagan (1995) formally proposed three ways to set the value of b: (i) $b = m_0/N$, when robustness is no concern, (ii) $b = N^{-1}\max\{m_0, N^{-1/2}\}$, when robustness is a serious concern, and (iii) $b = N^{-1}\max\{m_0, \log N\}$, as an intermediate option. Here m_0 denotes the smallest possible sample size permitting a comparison of M_0 to M_1. The minimal training sample requires $N_j = p + 1$, $N_i = 1$, $i = 1, \ldots, K$, $i \neq j$, for some $1 \leq j \leq K$ (since we need at least one observation in each group, to estimate μ_k $(k = 1, \ldots, K)$ plus p further observation in order to be able to estimate Σ).

3. Detection of Influential Observations

3.1. Diagnostic Measure

B_{01}^b is seen to be used as a measure for discriminatory information (i.e. the smaller B_{01}^b leads to the stronger evidence for the significant contribution to group separation), so that we may make use of it to develop a diagnostic measure for detecting influential observations in MDA. We see from the definition 1 that

$$P_b(D|M_i) = \frac{\int \pi(\theta_i|M_i) f(D_{(r)}|\theta_i, M_i) f(x_{(r)}|\theta_i, M_i) d\theta_i}{\int \pi(\theta_i|M_i) f(D_{(r)}|\theta_i, M_i)^b f(x_{(r)}|\theta_i, M_i)^b d\theta_i}$$

$$= P_b(D_{(r)}|M_i)f(x_{(r)}|D_{(r)}, M_i)^{(1-b)}, \text{ for } i = 0, 1,$$

where $D_{(r)}$ denotes all elements of the training sample D except the rth observation $X_{(r)}$, $r = 1, \ldots, N$.

Let $B^b_{01\backslash(r)}$ is the Bayes factor of M_0 relative to M_1 using all but rth observation in D, then

$$B^{(r)} = B^b_{01\backslash(r)}/B^b_{01} = \left(\frac{f(x_{(r)}|D_{(r)}, M_0)}{f(x_{(r)}|D_{(r)}, M_1)}\right)^{b-1},$$

where $-1 < (b-1) < 0$. This is $(b-1)$th power of the conditional predictive ordinate (CPO) ratio for model comparison that measures the contribution to the adequacy of the model M_1 attributable to the rth observation. See Pettit and Young (1990) and references therein.

Therefore, in order to find influential observations when using B^b_{01}, one has to compute the $B^{(r)} = B^b_{01\backslash(r)}/B^b_{01}$ for all observations in D, $r = 1, 2, \ldots, N$, and choose the rth observation, say $X_{(r^*)}$ having minimal value of $B^{(r)}$,

If $B^{(r^*)} > 1$, the r^*th observation contributes to the discriminatory information for MDA; If $B^{(r^*)} < 1$, it is outlier observation influential on the discriminatory information, where $B^{(r^*)} = \text{Min}\{B^{(r)}; r = 1, \ldots, N\}$. However, the value of $B^{(r^*)} > 1$ (or< 1) does not indicate whether deletion of r^*th observation is enough to change our beliefs form, say, supporting M_0 to supporting M_1. To assess whether deleting r^*th observation changes our beliefs we have to compare $B^{(r^*)}$ with B^b_{01}. We illustrate this in later examples.

3.2. Stopping Rule

An advantage of using $B^{(r)}$, besides its direct interpretation and simplicity of calculations compared to the error rates criterion, is that it naturally provides a stopping rule. Suppose the subset S_{N^*} of D achieves minimum B^b_{01} among all subsets of size N^* observations considered. Let $B^b_{01}(S_{N^*})$ be the corresponding value. Similarly for size $N^* - 1$ we have S_{N^*-1} with $B^b_{01}(S_{N^*-1})$. Either the ratio

$$\Delta(N^* - 1, N^*) = B^b_{01}(S_{N^*-1})/B^b_{01}(S_{N^*}) \tag{5}$$

or its logarithmic scale

$$C_r(N^*) = \log_{10} B^b_{01}(S_{N^*\backslash(r)}) - \log_{10} B^b_{01}(S_{N^*}), \text{ where } S_{N^*\backslash(r)} = S_{N^*-1},$$

gives a criterion for measuring the increase in the discriminatory information. When an additional influential observation is detected from the measure, the number of observations used in MDA are to be decreased from N^* to $N^* - 1$. So a stopping rule can be based on it by specifying a threshold value λ which is obtained by the Jeffreys' scale evidence of the Bayes factor in Table I. If $C_r(N^*) < \lambda$ when the subset $S_{N^*\setminus(r)}$ is used instead of S_{N^*} for MDA there is an increase evidence for M_1. Consequently $S_{N^*\setminus(r)}$ favors model M_1 more than S_{N^*} does, and hence its observations can be accounted for as having more discriminatory information for MDA than those of S_{N^*}. Similarly, if $C_r(N^*) \geq \lambda$ one can say that observations of $S_{N^*\setminus(r)}$ have less discriminatory information than those of S_{N^*}. According to the scale of evidence for assessing Bayes factors, shown Table I, we may take $\lambda = -0.5$. Thus $S_{N^*\setminus(r)}$ with $C_r(N^*) < -0.5$ might be thought of as significantly better training sample set than S_{N^*} in MDA. Formally, we have the following all subset approach: stop selection at step M if

$$C_r(M) = \log_{10}\{B^b_{01}(S_{M\setminus(r)})/B^b_{01}(S_M)\} \geq \lambda,$$

while

$$C_r(N^*) = \log_{10}\{B^b_{01}(S_{N^*\setminus(r)})/B^b_{01}(S_{N^*})\} < \lambda, \quad N^* = N-1,\ldots,M-1, \quad N-1 \geq M,$$

where $\lambda = -0.5$. Note that if $S_{N^*} \setminus S_{N^*\setminus(r)} = \{X_{(r)}\}$, $C_r(N^*) = \log_{10} B^{(r)}$.

Table I. Logarithmic scale of evidence for assessing Bayes Factor

Range	Evidence
$\log_{10} B^b_{01} > 0$	Supports M_0
$0 > \log_{10} B^b_{01} > -0.5$	Slight evidence against M_0
$-0.5 > \log_{10} B^b_{01} > -1.0$	Moderate evidence against M_0
$-1.0 > \log_{10} B^b_{01} > -2.0$	Strong evidence against M_0
$-2.0 > \log_{10} B^b_{01}$	Decisive evidence against M_0

4. Illustrative Examples

Two sampling experiments were carried out employing the diagnostic measure and the stopping rule. For the calculation of the diagnostic measure for each experiment, we assume that its robustness is no concern and hence we set $b = m_0/N$. As a tool for visual interpretation, a graphical method is presented in this section.

4.1. Remote Sensing Data

To examine the performance of the suggested detection method for influential observations, we use "Remote Sensing Data" obtained from SAS/STAT user's guide (1982 edition). The data on crops is collected for five crops and their observations are grouped into five groups: corn, soybean, cotton, sugar beets, and clover. Four measures called $x1 - x4$ make up the descriptive variables.

By means of program using SAS/IML, we analyzed this data. MDA using proportional (to sample sizes) prior probabilities for each group led to the actual error rate (AER) = 0.667. As advocated by Rencher (1995), we estimated AER by use of the cross-validation method. The discriminatory information $B_{01}^b = 2.9036 \times 10^{-7}$ indicates evidence for the significance discriminatory information in the training sample. However as can be seen from Figure 1a, there is the largest effect on B_{01}^b when we omit the 26th observation (first observation in clover group) yielding $C_{26}(36) = -2.017$. In other words, omitting the 26th observation changes Jeffreys' scale evidence of the significance of discriminatory information in MDA

Table 2. Remote Sensing Data

Corn				Soybean				Cotton				Sugarbeet				Clover			
$x1$	$x2$	$x3$	$x4$	$x1$	$x2$	$x3$	$x4$	$x1$	$x2$	$x3$	$x4$	$x1$	$x2$	$x3$	$x4$	$x1$	$x2$	$x3$	$x4$
16	27	31	33	20	23	23	25	31	32	33	34	22	23	25	42	12	45	32	54
15	23	30	30	24	24	25	32	29	24	26	28	25	25	24	26	24	58	25	34
16	27	27	26	21	25	23	24	34	32	28	45	34	25	16	52	87	54	61	21
18	20	25	23	27	45	24	12	26	25	23	24	54	23	21	54	51	31	31	16
15	15	31	32	12	13	15	42	53	48	75	26	25	43	32	15	96	48	54	62
15	32	32	15	22	32	31	43	34	35	25	78	26	54	2	54	31	31	11	11
12	15	16	73													56	13	13	71
																32	13	27	32
																36	26	54	32
																53	8	6	54
																32	32	62	16

According to our diagnostic measure, this observation can be taken as the most influential observation among the training sample. We see that omitting the influential observation leads to change of AER from .6667 to .5143. By use of the stopping rule in Subsection 3.2, we detected all the influential observations. They were found in the following order; $\{X_{(r)}; r = 26, 27, 36, 34, 23, 33, 31, 29, 22, 30\}$, where $X_{(r)}$ is rth observation in the data set (see Figure 1b). To save the space, figures are drawn only for the first stage (Figure 1a) and the final stage of the

stopping rule (Figure 1b), and they note $C_r = \log_{10} B^{(r)}$ for the influence of rth observation, $X_{(r)}$, $r = 1, \ldots, N^*$. Thus we may well regard there are at least 10 influential observations in the data set. Moreover, deletion of the 10 influential

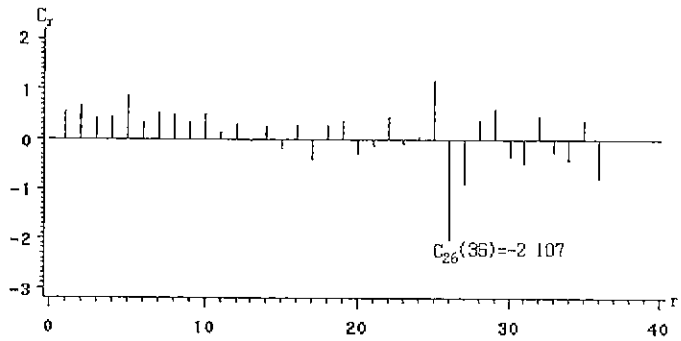

Figure 1a. C_r Plot for the First Stage of the Influential Selection, where $N^* = 36$.

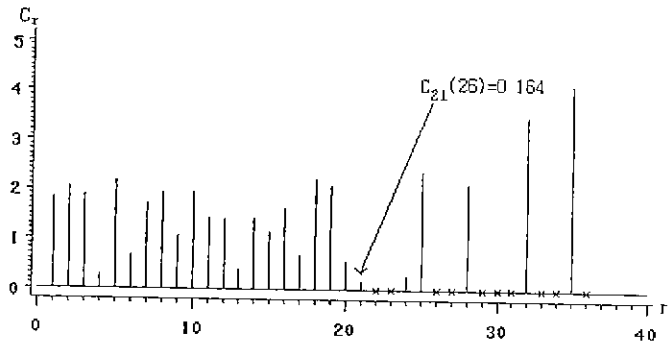

Figure 1b. C_r Plot for the Last Stage of the Selection Excluding 10 influential Observations, where $N^* = M = 26$.

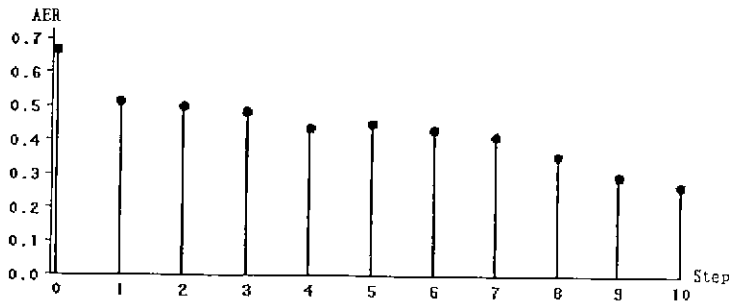

Figure 2. Change of AER Obtained from Each Step of the Stopping Rule.

observations reduces the value of $\log_{10} B_{01}^{b}$ from -6.537 to -29.817 substantially increasing the discriminatory information. Figure 2 depicts the change of AER obtained from deleting influential observations in each stage of the stopping rule. As noted by Figure 2, if we apply the stopping rule using the suggested diagnostic measure we can reduce AER from .667 to .269. Therefore, this example shows that the suggested diagnostic successfully reveals the influential observations and leads to decrease of AER, so that it may be used as a method for choosing observations to be excluded for improving the performance of MDA

4.2. Bisbey Data

This data set listed in Huberty (1994 p. 277) was constructed by Gerald D. Bisbey (1968). A sample of 153 students entering college was partitioned according to the college French course in which they enrolled. Thirty five ($N_1 = 35$) enrolled in the beginning level, $N_2 = 81$ in the intermediate level, and $N_3 = 37$ in the advanced level. The data consists of thirteen measures obtained on each of the 153 students.

The purpose of this example is to demonstrate performance of the suggested diagnostic measure of influential observations in MDA. By use of the diagnostic measure $C_r = \log_{10} B^{(r)}$, the stopping rule was again applied to the data set. It detects 20 influential observations in the following order of selection; $\{X_{(r)} : r = 32, 30, 106, 67, 125, 123, 57, 109, 9, 75, 86, 59, 19, 111, 40, 117,$ $124, 85, 48, 115\}$ (see Figure 3a and Figure 3b). The deletion of 20 influential observations reduces $\log_{10} B_{01}^{b}$ and AER from -87.6633 to -108.431 and from .1645 to .0526, respectively (see Figure 4). As the first illustrative example, this example

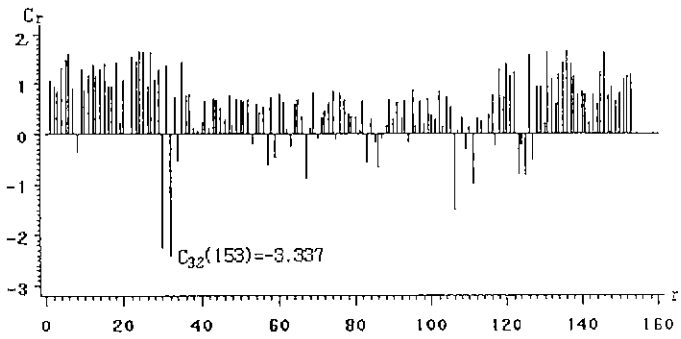

Figure 3a. C_r Plot for the First Stage of the Influential Selection, where $N^* = 153$.

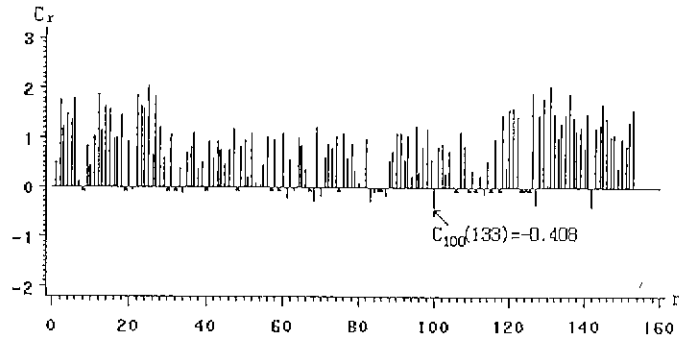

Figure 3b. C_r Plot for the Last Stage of the Selection, Excluding 20 influential Observations, where $N^* = M = 133$.

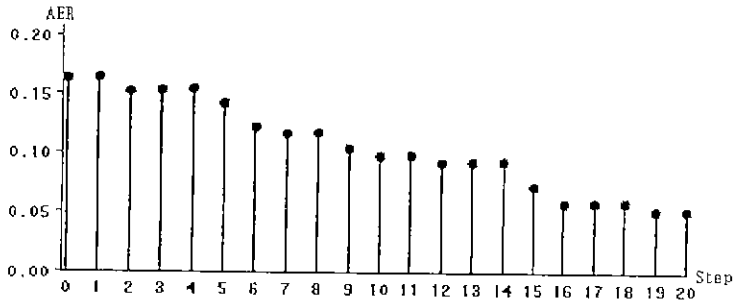

Figure 4. Change of AER Obtained from Each Step of the Stopping Rule.

also confirms us that the suggested diagnostic measure and the stopping rule is useful tools for detecting influential observations and improving performance of MDA. In this case, we may reuse 20 influential observations in the following way: Construct an optimal classification rule, and then allocate them into one of the three groups based upon the optimal rule.

5. Concluding Remarks

We proposed a new diagnostic measure for detecting single influential observation in MDA. When we apply the measure sequentially, it could also be useful for identifying multiple influential observations. The measure is developed from a Bayesian point of view using a default Bayes factor that can be taken as a Bayesian discriminant criterion in MDA. Based on the criterion, we suggests a diagnostic measure and a stopping rule for detecting observations that deteriorates

the discriminatory information. The diagnostic measure may be interpreted as incremental contribution to the discriminatory information in MDA attributable to a single observation. Illustrative examples in Section 4 confirm us that the suggested diagnostic measure and the stopping rule is useful tools for detecting influential observations and improving performance of MDA.

The proposed measure can be easily extended to detect multiple influential observations in block avoiding the masking problem (cf. Rousseeuw and Zomeren, 1990) and to detect influential observations in multiple discriminant analysis with heterogeneous covariance matrices. These problems are worthy to study and are left as a future subject of research.

References

Anderson, T. W. (1984). *An Introduction to Multivariate Statistical Analysis,* John Wiley, New York .

Berger, J. O. and Pericchi, L. (1996). The intrinsic Bayes factor for model selection and prediction, *Journal of the American Statistical Association,* Vol. 91, 109-122.

Berger, J. O. and Sellke, T. (1987). Testing a point null hypothesis: The reconciliability of P-values and evidence, *Journal of the American Statistical Association,* Vol. 82, 112-122.

Critchley, F. and Vitiello, C. (1991). The influence of observations on misclassification probability estimates in linear discriminant analysis, *Biometrika,* Vol. 78, 677-690.

DeGroot, M. H. (1970). *Optimal Statistical Decisions,* McGraw-Hill, New York.

Fung, W. K. (1992). Some diagnostic measures in discriminant analysis, *Statistics and Probability Letters,* Vol. 13, 279-285.

Fung, W. K. (1995). Diagnostics in linear discriminant analysis, *Journal of the American Statistical Association,* Vol. 90, 952-956.

Fung, W. K. (1999). Outlier diagnostics in several multivariate samples, *Statistician,* Vol. 48, Part 1, 73-84.

Hawkins, D.M. (1982). *Topics in Applied Multivariate Analysis,* Cambridge University Press, Cambridge.

Huberty, C. J. (1994). *Applied Discriminant Analysis,* John Wiley & Sons, New York.

Jeffreys, H. (1961). *Theory of Probability,* Oxford University Press.

Johnson, W. (1987). The detection of influential observations for allocation, separation, and the determination of probabilities in a Bayesian framework, *Journal of Business and Economic Statistics,* Vol. 5, 369-381.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors, *Journal of the American Statistical Association,* Vol. 90, 773-795.

Lee, P. M. (1988). *Bayesian Statistics: An Introduction,* John Wiley, New York.

Lempers, F. B. (1971). *Posterior Probabilities of Alternative Linear Models,* University Press, Rotterdam.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition,* John Wiley, New York.

O'Hagan, A. (1995). Fractional Bayes Factors for Model Comparisons, *Journal of the Royal Statistical Society,* B, Vol. 57, 99-138.

Pettit, L. I. and Young, K. D. S. (1990). Measuring the effect of observations on Bayes factors, *Biometrika,* Vol. 77, 455-466.

Rencher, A. C. (1995). *Methods of Multivariate Analysis,* John Wiley, New York.

Rousseeuw, P. J. and Zomeren V. (1990). Unmasking multivariate outliers and leverage points (with discussion), *Journal of the American Statistical Association,* Vol. 85, 633-651.

Spiegelhalter, D. J. and Smith, A. F. M. (1982). Bayes Factors for Linear and Log-linear Models with Vague Prior Information, *Journal of the Royal Statistical Society,* B, Vol. 44, 377-387.