The Korean Communications
in Statistics Vol. 3, No. 3, 1996
pp. 25-30

On Effect of Nonnormality on Size of Test for Dimensionality in Discriminant Analysis

Changha Hwang[1)]

Abstract

In discriminant analysis the procedures commonly used to estimate the dimensionality involve testing a sequence of dimensionality hypotheses. There is a problem with the size of the test since dimensionality hypotheses are tested sequentially and thus they are actually conditional tests. The focus of this paper is to investigate in asymptotic sense what happens to the sequential testing procedure if the assumption of normality does not hold.

1. Introduction

In discriminant analysis, the study of dimensionality is quite interesting since it determines the number of discriminant functions required to describe group differences. The procedures commonly used to estimate this dimensionality involve testing a sequence of dimensionality hypotheses. These hypotheses are tested sequentially and thus they are actually conditional tests; that is, we test H_k after we have tested and rejected the hypotheses H_0, $H_1, \cdots, H_{k-1}$ in sequence. There is a problem with the size of the test since successive tests are not independent. Hwang(1995b) showed that the size of test is not affected asymptotically under the normality. The focus of this paper is to investigate in asymptotic sense "How is the size of the test affected under nonnormality by viewing this sequence of tests as conditional tests?".

2. Main Result

Let $\boldsymbol{y}_{i1}, \cdots, \boldsymbol{y}_{iq_i}$ $(i=1,\cdots,p)$ be i.i.d. $m\times 1$ absolutely continuous random vectors with mean $\boldsymbol{\mu}_i$, covariance matrix $\boldsymbol{\Sigma}$ and finite fourth moments. Suppose that the samples are independent across populations. Let $\overline{\boldsymbol{y}}_i$ be the sample mean of the q_i observations in the i

1) Assistant Professor, Dept. of Statistics, Catholic University of Taegu-Hyosung, Kyungbuk, 713-702, Korea.

th sample and $\bar{y}$ be the sample mean of all n observations, $(n = \sum_{i=1}^{p} q_i)$. Then matrices A and B are defined as

$$A = \sum_{i=1}^{p} q_i(\bar{y}_i - \bar{y})(\bar{y}_i - \bar{y})' \text{ and } B = \sum_{i=1}^{p}\sum_{j=1}^{q_i}(y_{ij} - \bar{y}_i)(y_{ij} - \bar{y}_i)'.$$

The matrix Ω is defined as $\Omega = \Sigma^{-1}\sum_{i=1}^{p} q_i(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'$, where $\bar{\mu} = \frac{1}{n}\sum_{i=1}^{p} q_i \mu_i$. From now on, we will assume that $p \geq m+1$ so that AB^{-1} has m nonzero eigenvalues $f_1 > \cdots > f_m > 0$. For the asymptotic theory there is no loss of generality in assuming that Ω is the diagonal matrix defined by $\Omega = diag\{w_1, \cdots, w_m\}$, $\Omega = n_2\Theta$, and $\Sigma = I_m$ where $n_2 = n - p$ and Θ is the fixed matrix defined by $\Theta = diag\{\theta_1, \cdots, \theta_m\}$. This means that we consider the case where A, B, Ω, and Σ are already transformed to canonical form. Thus, the dimensionality is the rank of Ω.

In practice, to determine the number of useful discriminant functions we test the sequence of dimensionality hypotheses,

$$H_k:\ \theta_{k+1} = \cdots = \theta_m = 0\ (\theta_k > 0),\ k = 0, 1, \cdots, m-1.$$

By testing these hypotheses sequentially they are actually conditional tests. We test H_k given we have tested and rejected H_0, H_1, $\cdots$, H_{k-1}, keeping in mind the effect on the significance level(the size of test). The likelihood ratio test statistic for H_k is given by

$$T_k = n_2 \sum_{i=k+1}^{m} \log(1 + f_i)$$

where $f_1 > \cdots > f_m > 0$ are the eigenvalues of AB^{-1}. For nonnormal populations, the asymptotic distribution of T_k is $\chi^2_{(m-k)(n_1-k)}$ when H_k is true. See for details Hwang(1994).

For our purpose, we need the following asymptotic expansion of test statistic T_k:

$$T_k = n_2 \sum_{i=k+1}^{m} \log(1 + \theta_i) + \sqrt{n_2}C + D + O_p(n_2^{-\frac{1}{2}}),$$

where

$$C = \sum_{i=k+1}^{m} \frac{E_{ii}(n) - \theta_i U_{ii}(n)}{1 + \theta_i},$$

$$D = \sum_{i=k+1}^{m} \frac{F_{ii}(n)}{(1 + \theta_i)} - \sum_{i=1}^{k}\sum_{j=k+1}^{m} \frac{E_{ij}(n)^2}{(1 + \theta_j)(\theta_i - \theta_j)}$$

$$+ \sum_{i=1}^{k}\sum_{j=k+1}^{m} \frac{4\theta_j E_{ij}(n)U_{ij}(n)}{(1 + \theta_i)(1 + \theta_j)(\theta_i - \theta_j)} - \sum_{i=k+1}^{m}\sum_{j=k+1}^{m} \frac{E_{ij}(n)U_{ij}(n)}{(1 + \theta_i)(1 + \theta_j)}$$

$$+ \sum_{i=1}^{k} \sum_{j=k+1}^{m} \frac{\theta_j(\theta_j - 2\theta_i - \theta_i^2)\, U_{ij}(n)^2}{(1+\theta_i)(1+\theta_j)(\theta_i - \theta_j)} + \sum_{i=k+1}^{m} \sum_{j=k+1}^{m} \frac{\theta_i(\theta_j + 2)\, U_{ij}(n)^2}{2(1+\theta_i)(1+\theta_j)}$$

$$- \sum_{i=k+1}^{m} \sum_{j=k+1}^{m} \frac{E_{ij}(n)^2}{2(1+\theta_i)(1+\theta_j)}.$$

Furthermore, $E_{ij}(n)$, $F_{ij}(n)$, and $U_{ij}(n)$ are the ijth element of matrices $E(n)$, $F(n)$, and $U(n)$ defined as follows:

$$E(n) = \frac{1}{\sqrt{n_2}} \sum_{i=1}^{p} q_i [(\bar{\varepsilon}_i - \bar{\varepsilon})(\mu_i - \bar{\mu})' + (\mu_i - \bar{\mu})(\bar{\varepsilon}_i - \bar{\varepsilon})'],$$

$$F(n) = \sum_{i=1}^{p} q_i (\bar{\varepsilon}_i - \bar{\varepsilon})(\bar{\varepsilon}_i - \bar{\varepsilon})',$$

$$U(n) = \frac{1}{\sqrt{n_2}} \sum_{i=1}^{p} \sum_{j=1}^{q_i} [(\varepsilon_{ij} - \bar{\varepsilon}_i)(\varepsilon_{ij} - \bar{\varepsilon}_i)' - I_m],$$

where $\bar{y}_i = \mu_i + \bar{\varepsilon}_i$, $\bar{y} = \bar{\mu} + \bar{\varepsilon}$, $\bar{\varepsilon}_i = \frac{1}{q_i}\sum_{j=1}^{q_i} \varepsilon_{ij}$, and $\bar{\varepsilon} = \frac{1}{n}\sum_{i=1}^{p} q_i \bar{\varepsilon}_i$. See for details Hwang(1995a).

Theorem 1 For each $i = 1, \cdots, p$, let $y_{ij}: m \times 1$, $j = 1, \cdots, q_i$ be a sequence of i.i.d. random vectors drawn from m multivariate elliptical populations with parameter μ_i, covariance matrix I_m, and finite fourth moments. Suppose that the p sequences are independent and put

$$T_k = n_2 \sum_{i=k+1}^{m} \log(1 + f_i)$$

$$V_k = \frac{1}{\sqrt{n_2}} [T_k - n_2 \sum_{i=k+1}^{m} \log(1 + \theta_i)]$$

Then under H_k, T_k is asymptotically independent of V_j, $j = 0, 1, \cdots, k-1$.

Proof From the expansions of $T_0, T_1, \cdots, T_k$ under H_k we form two subvectors z_1 and z_2, where z_1 contains the $\sqrt{q_i}\, \bar{\varepsilon}_{i\cdot s}$ variables which make up T_k and z_2 contains the $\sqrt{q_i}\, \bar{\varepsilon}_{i\cdot t}$ and $\frac{1}{\sqrt{q_i}} \sum_{j=1}^{q_i} (\varepsilon_{ijt}^2 - 1)$ variables which make up $V_0, V_1, \cdots, V_{k-1}$. Specifically,

$$z_1 = (\sqrt{q_i}\, \bar{\varepsilon}_{i\cdot k+1}, \cdots, \sqrt{q_i}\, \bar{\varepsilon}_{i\cdot m})'$$

$$z_2 = (\sqrt{q_i}\, \bar{\varepsilon}_{i\cdot 1}, \cdots, \sqrt{q_i}\, \bar{\varepsilon}_{i\cdot k};\ \frac{1}{\sqrt{q_i}} \sum_{j=1}^{q_i} (\varepsilon_{ij1}^2 - 1), \cdots, \frac{1}{\sqrt{q_i}} \sum_{j=1}^{q_i} (\varepsilon_{ijk}^2 - 1))'$$

and define $z = (z_1', z_2')'$. Here, ε_{ij}, $\bar{\varepsilon}_i$ and $\bar{\varepsilon}$ are denoted by

$$\varepsilon_{ij} = (\varepsilon_{ij1}, \cdots, \varepsilon_{ijm})',\ \bar{\varepsilon}_i = (\bar{\varepsilon}_{i.1}, \cdots, \bar{\varepsilon}_{i.m})' \text{ and } \bar{\varepsilon} = (\bar{\varepsilon}_{..1}, \cdots, \bar{\varepsilon}_{..m})',$$

where $\bar{\varepsilon}_{i.r} = \frac{1}{q_i}\sum_{j=1}^{q_i}\varepsilon_{ijr}$ and $\bar{\varepsilon}_{..r} = \frac{1}{n}\sum_{i=1}^{p}\sum_{j=1}^{q_i}\varepsilon_{ijr}$, $r=1,\cdots,m$. By the Multivariate Central Limit Theorem z converges in distribution to multivariate normal with mean $\mathbf{0}$. The elements of asymptotic covariance matrix can be computed as follows: For $s=k+1,\cdots,m$, $t=1,\cdots,k$, as $n\to\infty$,

$$\begin{aligned} Cov(\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot s}, \sqrt{q_i}\,\bar{\varepsilon}_{i\cdot t}) &= E[(\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot s})(\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot t})] \\ &\to 0, \end{aligned}$$

since $\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot s}$ and $\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot t}$ are asymptotically independent. Also,

$$\begin{aligned} Cov\left(\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot s}, \frac{1}{\sqrt{q_i}}\sum_{j=1}^{q_i}(\varepsilon_{ijt}^2-1)\right) &= E\left[(\sqrt{q_i}\,\bar{\varepsilon}_{i\cdot s})\left(\frac{1}{\sqrt{q_i}}\sum_{j=1}^{q_i}(\varepsilon_{ijt}^2-1)\right)\right] \\ &= E[\varepsilon_{i1s}\varepsilon_{i1t}^2] = 0 \end{aligned} \qquad (1)$$

because for elliptical distributions all third moments are zero (see, for example, Gang(1987)). These covariance expressions show that the elements of z_1 are asymptotically independent of the elements of z_2 for elliptical distributions. Thus, under H_k, the test statistic T_k is asymptotically independent of the statistics V_j, $j=0,\cdots,k-1$. ∎

Therefore, this result agrees with the result for multivariate normal distributions. From (1) we see that the elements of z_1 are generally not asymptotically independent of the elements of z_2. Thus, under H_k, the test statistic T_k is generally not asymptotically independent of the statistics V_j, $j=0,\cdots,k-1$. It is shown that this result is sensitive to certain departures from normality.

3. Simulation Study and Conclusion

A Monte Carlo experiment was carried out to see how inferences regarding the sequential testing procedure based on the assumption of multivariate normality are affected if this assumption is violated. In particular, suppose we are sampling from an elliptical t-distribution on 5 degrees of freedom. Recall that the asymptotic distribution of $T_k = n_2\sum_{i=k+1}^{m}\log(1+f_i)$ is $\chi^2_{(n_1-k)(m-k)}$ for nonnormal populations. The statistic T_k was used in the study. The study consisted of generating 500 samples of size $n_2 = 50, 100, 200$ of an 4-variate elliptical t-distribution on 5 degrees of freedom for 6 populations with parameters $\mu_i (i=1,\cdots,6)$ and $V = \frac{3}{5}I_4$. These samples can be generated using the following relationship:

$$y_{ij} = \mu_i + Z^{-\frac{1}{2}}(4V)x,$$

where $x \sim N_4(0, I_m)$ and $Z \sim \chi^2_5$. For further details of simulations see Hwang(1994a). Generation of the samples, computation of the sample eigenvalue and the analysis were conducted using SAS and SAS/IML.

Table 1 and 2 present the observed unconditional and conditional significance levels, respectively. As n_2 increases, the observed conditional and unconditional significance levels become closer to each other. This result agrees with the result for multivariate normal distributions. To conclude, we see for multivariate elliptical populations the size of the test is not affected asymptotically by viewing this sequence of tests as conditional tests but this result is sensitive to certain departures from normality.

Table 1: **Unconditional Significance Levels for elliptical t(5)**,(m=4,p=6)

θ_1	θ_2	θ_3	θ_4	V	$T^{.05}$			$T^{.10}$		
					$n_2=50$	$n_2=100$	$n_2=200$	$n_2=50$	$n_2=100$	$n_2=200$
0.2	0	0	0	1	0.020	0.038	0.060	0.052	0.086	0.116
0.8	0	0	0	1	0.038	0.040	0.056	0.090	0.110	0.114
6	0	0	0	1	0.042	0.044	0.052	0.094	0.108	0.108
0.4	0.2	0	0	2	0.020	0.052	0.062	0.050	0.100	0.114
0.8	0.4	0	0	2	0.026	0.062	0.068	0.084	0.116	0.118
6	2	0	0	2	0.042	0.062	0.062	0.110	0.118	0.124
0.4	0.2	0.1	0	3	0.004	0.004	0.006	0.012	0.014	0.032
2	1	0.8	0	3	0.042	0.052	0.052	0.090	0.120	0.118
6	4	2	0	3	0.060	0.060	0.066	0.124	0.126	0.114

Table 2: **Conditional Significance Levels for elliptical t(5)**,(m=4,p=6)

θ_1	θ_2	θ_3	θ_4	V	$T^{.05}$			$T^{.10}$		
					$n_2=50$	$n_2=100$	$n_2=200$	$n_2=50$	$n_2=100$	$n_2=200$
0.2	0	0	0	1	0.062	0.053	0.061	0.106	0.103	0.117
0.8	0	0	0	1	0.041	0.040	0.056	0.092	0.110	0.114
6	0	0	0	1	0.042	0.044	0.052	0.094	0.108	0.108
0.4	0.2	0	0	2	0.058	0.066	0.063	0.150	0.113	0.115
0.8	0.4	0	0	2	0.036	0.063	0.068	0.099	0.117	0.118
6	2	0	0	2	0.042	0.062	0.062	0.110	0.118	0.124
0.4	0.2	0.1	0	3	0.250	0.222	0.059	0.250	0.214	0.174
2	1	0.8	0	3	0.067	0.054	0.052	0.118	0.122	0.118
6	4	2	0	3	0.060	0.060	0.066	0.124	0.126	0.114

References

[1] Gang, L. (1987). Moments of a random vector and its quadratic forms. *J. Statist. Appl. Prob.,* 2, 219-229. [Reprinted in Statistical Inference in Elliptically Contoured and Related Distributions(Fang, K. and Anderson T.W., ed.), Allerton Press, 1990, 433-440.]

[2] Hwang, C. (1991). Model Selection Methods in Discriminant Analysis. Ph.D Thesis, Univ. of Michigan, Ann Arbor, Michigan.

[3] Hwang, C. (1994a). On estimating the dimensionality in discriminant analysis. *Communications in Statistics:Theory and Methods,* 23, 2197-2215.

[4] Hwang, C. (1994b). Characterization of the asymptotic distributions of certain eigenvalues in a general setting. *Journal of the Korean Statistical Society,* 23, 13-32.

[5] Hwang, C. (1995a). A note on the asymptotic distributions of dimensionality Estimators in discriminant analysis. *The Korean Communications in Statistics,* 2, 320-329.

[6] Hwang, C. (1995b). Size of test for dimensionality in discriminant analysis. *Journal of Statistical Theory and Methods,* 6, 9-15.

[7] Seo, T., Kanda, T. and Fujikoshi, Y. (1993). The effects of nonnormality on tests for dimensionality in canonical correlation and MANOVA models. Technical Report No. 93-9, Hiroshima University.

[8] Siotani, M., Hayakawa, T., and Fujikoshi, Y. (1985). Modern Multivariate Statistical Analysis: A Graduate Course and Handbook. American Sciences Press, Inc.