Journal of Statistical
Theory & Methods
1995, Vol.6, pp 9~15

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 "How is the size of the test affected by viewing this sequence of tests as conditional tests?".

Keyword: dimensionality, size of test.

1. Introduction

In discriminant analysis, the study of dimensionality of the hyperplane 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. The focus of this paper is to investgate for normal populations "How is the size of the test affected by viewing this sequence of tests as conditional tests?".

Let $y_{i1}, \cdots, y_{iq_i} (i = 1, \cdots, p)$ be independent $N_m(\mu_i, \Sigma)$ random vectors. Suppose that the samples are independent across populations. Also let $\bar{y}_{ij}$ be the sample mean of the q_i observations in the i^{th} sample $(i = 1, \cdots, p)$ 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

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

[1] Dept. of Statistics, Catholic University of Taegu-Hyosung,Kyungbuk 712-702.

The distributions of A and B are

$$A \sim W_m(p-1, \Sigma, \Omega) \text{ and } B \sim W_m(n-p, \Sigma),$$

where $\Omega = \Sigma^{-1} \sum_{i=1}^{p} q_i(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})'$ and $\bar{\mu} = \frac{1}{n} \sum_{i=1}^{p} q_i \mu_i$. 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 $\Omega = n_2\Theta$, and $\Sigma = I_m$ where $n_2 = n-p$ and Θ is the fixed matrix defined by $\Theta = diag\{\theta_1, \cdots, \theta_m\}$. In multiple discriminant analysis the dimensionality is, in fact, the rank of Ω.

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

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

By testing these hypotheses sequentially they are actually conditional tests. We test H_k given we have tested and rejected $H_0, \cdots, H_{k-1}$, keeping in mind the effect on the significance level(the size of test). For example, suppose the smallest $m-1$ population roots are zero, that is, suppose the null hypothesis $H_1 : \theta_2 = \cdots = \theta_m = 0 \ (\theta_1 > 0)$ is true. Put $n_1 = p-1$. Then, under H_1 we have that

$$P_{H_1}[reject\ H_1] = p_{H_1}[T_1 > c_{f_1}(\alpha)] = \alpha$$

asymptotically, where $c_{f_1}(\alpha)$ is the upper $100\alpha\%$ point of $\chi^2_{(m-1)(n_1-1)}$. This is the unconditional level of significance. But we are testing H_1 because we rejected the previous null hypothesis, H_0, that all the population eigenvalues are zero. How can we compare the conditional level of significance, that is,

$$P_{H_1} = [reject\ H_1 | reject\ H_0] = P_{H_0}[T_1 > c_{f_1}(\alpha) | T_0 > c_{f_0}(\alpha)]$$

to the unconditional level of significance, α, asymptotically? Note $c_{f_0}(\alpha)$ is the upper $100\alpha\%$ point of $\chi^2_{(mn_1)}$.

The likelihood ratio test rejects H_k for small values of the statistic $W_k = \prod_{i=k+1}^{m}(1+f_i)^{-1}$ Again assume that $n_1 \geq m$ and $n_2 \geq m$, where $n_1 = p-1$ and $n_2 = n-p$. Put $T_k = -n_2 \log W_k$ so that the asymptotic distribution of T_k is $\chi^2_{(m-k)(n_1-k)}$ when H_k is true.

2. Main Result

Using an idea in Siotani *et al*(1985),pg. 463 we can write $\frac{1}{n_2} A$ as

$$\frac{1}{n_2}A = \Theta + \frac{1}{\sqrt{n_2}}Z + \frac{1}{n_2}YY' \tag{1}$$

where $Y=[y_1,\cdots,y_{n_1}]$, y_i's are independently distributed as $N_m(0,I_m)$, $Y=[Y_1,Y_2]$, $Y_1:m\times m$ and $Z=Y_1\Theta^{\frac{1}{2}}+\Theta^{\frac{1}{2}}Y_1'$. Note that $y_1,\cdots,y_{n_1}$ are different from random vectors $y_{i1},\cdots,y_{iq_i}(i=1,\cdots,p)$ given before. Let

$$\frac{1}{n_2}B = I_m + \frac{1}{\sqrt{n_2}}U \tag{2}$$

Then the limiting distribution of $U=(U_{ij})$ is normal with mean zero and covariance matrix with elements $Cov(u_{ij},u_{kl})=\delta_{ik}\delta_{jl}+\delta_{il}\delta_{jk}$, where δ_{ij} is the Kronecker delta. One can obtain an asymptotic expansion for f_i when θ_i is simple. This expansion is given by

$$f_i = \theta_i + \frac{1}{\sqrt{n_2}}c_{ii} + \frac{1}{n_2}\{(YY')_{ii} + \sum_{j\neq i}^{m}\frac{c_{ji}}{\theta_i-\theta_j} - c_{ii}u_{ii}\} + O_p(n_2^{-\frac{3}{2}}), \tag{3}$$

where y_{ij} and u_{ij} are the j^{th} and the ij^{th} element of y_i and U, respectively, and $c_{ji}=z_{ji}-\theta_i u_{ji}=\sqrt{\theta_i}\,y_{ji}+\sqrt{\theta_j}\,y_{ij}-\theta_i u_{ji}$. See for details Siotani *et al*(1985) or Sugiura(1976).

For our purpose we derive an asymptotic expansion for T_k in the following way using the idea given in Muirhead and Waternaux (1980). The test criterion T_k is rewritten as

$$T_k = n_2[\log|I+AB^{-1}| - \sum_{i=1}^{k}\log(1+f_i)]. \tag{4}$$

For the asymptotic theory we assume the population eigenvalue θ_i is simple. Substituting (1), (2) and (3) in the expression (4) for T_k, we can show,after straightforward but lengthy algebraic manipulation, that T_k has the following expansion:

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

where

$$C=\sum_{i=k+1}^{m}\frac{2\sqrt{\theta_i}y_{ii}-\theta_i u_{ii}}{1+\theta_i},$$

$$D=\sum_{i=k+1}^{m}\sum_{j=1}^{n_1}\frac{y_{ij}^2}{1+\theta_i}-\sum_{i=k+1}^{m}\sum_{j=1}^{k}\frac{\theta_j y_{ij}^2}{(1+\theta_i)(\theta_j-\theta_i)}$$
$$+\sum_{i=k+1}^{m}\sum_{j=1}^{k}\frac{\theta_i y_{ji}^2}{(1+\theta_i)(\theta_i-\theta_j)}+\sum_{i=k+1}^{m}\sum_{j=1}^{k}\frac{2\sqrt{\theta_i\theta_j}y_{ij}y_{ji}}{(1+\theta_i)(\theta_i-\theta_j)}$$
$$-\sum_{i=k+1}^{m}\sum_{j=1}^{k}\frac{2\theta_i\sqrt{\theta_j}y_{ij}u_{ij}}{(1+\theta_i)(\theta_i-\theta_j)}+\sum_{i=k+1}^{m}\sum_{j=1}^{k}\frac{\sqrt{\theta_i}(2\theta_i-\theta_j+\theta_i\theta_j)y_{ji}u_{ji}}{(1+\theta_i)(1+\theta_j)(\theta_i-\theta_j)}$$
$$+\sum_{i=k+1}^{m}\sum_{j=1}^{k}\frac{\theta_i(\theta_i-2\theta_j+\theta_i\theta_j)u_{ji}^2}{(1+\theta_i)(1+\theta_j)(\theta_i-\theta_j)}-\sum_{i=k+1}^{m}\sum_{j=k+1}^{m}\frac{2\sqrt{\theta_i}y_{ji}u_{ji}}{(1+\theta_i)(1+\theta_j)}.$$

Theorem 1. Let $f_i\ (i=1,\cdots,m)$ be the eigenvalues of AB^{-1}, where $A\sim W_m(n_1,I_m,n_2\Theta)$ and $B\sim W_m(n_2,I_m)$. Let

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

When the null hypothesis H_k is true, 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 the two subvectors z_1 and z_2, where z_1 contains the y_{ij} variables which make up T_k and z_2 contains the y_{ji} and u_{ii} variables which make up $V_0,\cdots,V_{k-1}$.

$$z_1=(y_{k+1,k+1},\cdots,y_{k+1,n_1};y_{k+2,k+1},\cdots,y_{k+2,n_1};\cdots;y_{m,k+1},\cdots,y_{m,n_1})$$
$$z_2=(y_{11},u_{11};y_{22},u_{22};\cdots;y_{kk},u_{kk}).$$

Note that the limiting distribution (under normality) of $U=(u_{ij})$ is normal with mean zero and $E(u_{ii}^2)=2$, and that $Y=(y_{ij})$ and U are independently distributed, y_{ij}'s being independent $N(0,1)$ variables. Using the Multivariate Centeral Limit Theorem we have

$$z=(z_1',z_2')'\xrightarrow{L}N(0,\Sigma)$$

where

$$\Sigma=\begin{pmatrix}I_{(m-k)(n_1-k)} & 0\\ 0 & \Gamma\end{pmatrix}$$

and $\Gamma = diag(\Gamma_1, \cdots, \Gamma_k)$ with

$$\Gamma_i = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}.$$

Hence, we show that T_k and V_j, $j = 0,1,\cdots,k-1$ are asymptotically independent.

3. Monte Carlo Study

From Theorem 1 we realize that the size of test is not affected asymptotically under the normality. A Monte Carlo Study was conducted to investigate this. The study consisted of generating 500 values of a noncentral *Wishart* matrix A and a central *Wishart* matrix B with $m = 4, p = 6 (i.e., n_1 = 5), n_2 = 50,100,200$. Generation of the samples, computation of the sample eigenvalues and the analysis were conducted using SAS and SAS/IML.

The null hypothesis $H_1: \theta_2 = \theta_3 = \theta_4 = 0$ is tested with varying values of the largest eigenvalue θ_1 in the first stage of the study. The results of this study are presented in Table 1. For each value of n_2 and θ_1, the number of times out of 500 that the null hypothesis H_1 is rejected with the likelihood ratio test statistic $T_1 = -n_2 \log W_1$ is shown together with the corresponding observed unconditional significance level in parentheses. We compute the number of times that T_1 exceeds the upper $100\alpha\%$ to investigate how the size of unconditional test changes. From Table 1 we see as either n_2 or θ_1 increases the observed (unconditional) significance level approaches the appropriate nominal level. The problem arises since the observed significance level is much smaller than the nominal level when n_2 is small and θ_1 is close to zero. Overall these results for the unconditional level of significance agree with the distributional theory.

Next we consider only those cases for which the null hypothesis $H_0: \theta_1 = \theta_2 = \theta_3 = \theta_4 = 0$ was rejected with the likelihood ratio test statistic T_0. The proportion of these cases that reject H_1 gives the conditional level of significance. The observed conditional significance levels are summarized in Table 2. The unconditional level of the test(see Table 1) is smaller than nominal level when n_2 is small and θ_1 is close to zero; however as seen from Table 1 the conditional level of the test(see Table 2) is larger than the nominal level. Also as either n_2 or θ_1 increases the two observed significance levels approach the appropriate nominal level. Table 1 and 2 show that asymptotically the size of the test is not affected by conditioning on the event "reject H_0".

The simulation studies for hypotheses H_2, H_3 can be conducted using the same basic argument. Table 1 and 2 summarize simulation results for those cases.

Table 1. Unconditional Significance Levels under normal sampling,(m=4,p=6)

θ_1	θ_2	θ_3	θ_4	K	$T^{.05}$ $n_2=50$	$n_2=100$	$n_3=200$	$T^{.10}$ $n_2=50$	$n_2=100$	$n_3=200$
0.2	0	0	0	1	0.022	0.030	0.052	0.056	0.092	0.092
0.8	0	0	0	1	0.040	0.050	0.058	0.086	0.108	0.108
6	0	0	0	1	0.050	0.050	0.066	0.080	0.108	0.108
0.4	0.2	0	0	2	0.024	0.052	0.056	0.042	0.098	0.100
0.8	0.4	0	0	2	0.040	0.064	0.058	0.076	0.112	0.102
6	2	0	0	2	0.044	0.068	0.052	0.088	0.104	0.104
0.4	0.2	0.1	0	3	0.014	0.026	0.058	0.040	0.062	0.110
2	1	0.8	0	3	0.052	0.050	0.064	0.082	0.100	0.112
6	4	2	0	3	0.056	0.056	0.064	0.094	0.102	0.111

Table 2. Conditional Significance Levels under normal ampling,(m=4,p=6)

θ_1	θ_2	θ_3	θ_4	K	$T^{.05}$ $n_2=50$	$n_2=100$	$n_2=200$	$T^{.10}$ $n_2=50$	$n_2=100$	$n_2=200$
0.2	0	0	0	1	0.073	0.042	0.053	0.126	0.109	0.093
0.8	0	0	0	1	0.042	0.050	0.058	0.089	0.108	0.108
6	0	0	0	1	0.050	0.050	0.066	0.080	0.108	0.118
0.4	0.2	0	0	2	0.086	0.067	0.056	0.105	0.112	0.100
0.8	0.4	0	0	2	0.055	0.064	0.058	0.090	0.112	0.102
6	2	0	0	2	0.044	0.068	0.060	0.088	0.104	0.104
0.4	0.2	0.1	0	3	0.159	0.051	0.064	0.185	0.096	0.114
2	1	0.8	0	3	0.052	0.050	0.064	0.082	0.100	0.112
6	4	2	0	3	0.056	0.056	0.064	0.094	0.102	0.111

References

Hwang,C.(1991). Model selection methods in discriminant analysis. Ph.D Thesis, Univ. of Michigan, Ann Arbor, Michigan.

Hwang,C.(1994). On estimating the dimensionality in discriminant analysis. *Communications in Statistics* **23** 2197-2215.

Muirhead, R.J. and Waternaux, C.M.(1980). Asymptotic distributions in canonical correlation analysis and other multivariate procedures for nonnormal populations. *Biometrika* **67** 31-43.

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.

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

Sugiura, N.(1976). Asymptotic expansions of the distributions of the latent roots and the latent vectors of the Wishart and multivariate F matrices. *J. Multi. Anal.* **6** 500-525.