The Korean Communications
in Statistics Vol. 12 No. 2, 2005
pp. 473-481

Likelihood Ratio Criterion for Testing Sphericity from a Multivariate Normal Sample with 2-step Monotone Missing Data Pattern

Byungjin Choi[1)]

Abstract

The testing problem for sphericity structure of the covariance matrix in a multivariate normal distribution is introduced when there is a sample with 2-step monotone missing data pattern. The maximum likelihood method is described to estimate the parameters on the basis of the sample. Using these estimates, the likelihood ratio criterion for testing sphericity is derived.

Keywords : 2-step monotone missing data pattern, maximum likelihood estimation, sphericity, likelihood ratio criterion

1. Introduction

In applications, a researcher gathers observations on many correlated variables through the proper process to analyze the models considered in the study. In certain cases, however, some observations may take missing components due to the unknown factors such as, for instance, incorrect measurements and misrecording, etc. Since all of the standard statistical techniques used for analysis can be applied to complete observations, the slightly modified techniques are required to handle incomplete observations.

Toward this end, several methods such as, for example, the direct modeling and the conditional modeling have been suggested to make the statistical analysis of incomplete observations and a number of analysts have much contributed to the estimation and testing problems related to various statistical models developed for analyzing incomplete observations. See, for instance, Lord(1955), Anderson(1957), and Eaton and Kariya(1983). Bhargava(1975) discussed three inference problems on the mean vector of a multivariate normal distribution when some observations have a special missing pattern called monotone or monotone missing

1) Full-time Lecturer, Department of Applied Information Statistics, Kyonggi University, Suwon, Gyeonggi-do 443-760, Korea
E-mail : bjchoi92@kyonggi.ac.kr

data pattern and thereafter, statistical methods for handling these observations have been introduced and investigated by Little and Rubin(1987), Anderson and Olkin(1985), Jinadasa and Tracy(1992), and Provost(1990).

Sphericity plays an important role in the classical analysis of linear models such as the analysis of variance of repeated measurement data. Also, it is of direct interest in the field of applications such as animal navigation and astronomy. See Mardia and Jupp(1999) and references therein. Sphericity implies that the covariance matrix of a random variable vector is proportional to the identity matrix, this is, $\Sigma=\sigma^2 I$, where σ^2 is an unknown constant. Mauchly(1940) obtained the likelihood ratio criterion and its moments under the null hypothesis, $H_0 : \Sigma = \sigma^2 I$, and thereafter, many researchers have extensively investigated the testing problem for sphericity. See John(1972), Sugira(1972), Gupta(1977), Nagar et al.(1991), and Amey and Gupta(1992).

In this paper, we consider the testing problem for sphericity when there is a multivariate normal sample with 2-step monotone missing data pattern. In sections 2 and 3, we introduce the structure of this sample and describe the maximum likelihood estimation of the parameters based on the sample. In section 4, a likelihood ratio criterion for testing sphericity is derived using the estimates presented in the previous section. Section 5 gives a numerical example to implement the testing procedure.

2. Sample with 2-step Monotone Missing Data Pattern

Let $\boldsymbol{x}_k=(x_1, \cdots, x_{p_k})^t$, $k=0,1$, denote a random vector such that $\boldsymbol{x}_k \sim N_{p_k}(\boldsymbol{\mu}_k, \Sigma_{kk})$, where $\boldsymbol{\mu}_k=(\mu_1, \cdots, \mu_{p_k})^t$, Σ_{kk} is the covariance matrix of $\boldsymbol{x}_k$, and $0<p_1<p_0=p$. Let $\boldsymbol{x}_{i,0}=(x_{i1}, \cdots, x_{i,p})^t$, $i=1,\cdots,n_1$, denote an observation on $\boldsymbol{x}_0$ and let $\boldsymbol{x}_{i,1}=(x_{i1}, \cdots, x_{i,p_1})^t$, $i=n_1+1,\cdots,n$, denote an observation on $\boldsymbol{x}_1$, where n is overall sample size. Then the observations from the mutivariate normal distribution take the form of

$$X=\begin{pmatrix} x_{11} & \cdots & x_{1,p_1} & \cdots & x_{1,p} \\ \vdots & & \vdots & & \vdots \\ x_{n_1,1} & \cdots & x_{n_1,p_1} & \cdots & x_{n_1,p} \\ x_{n_1+1,1} & \cdots & x_{n_1+1,p_1} & & \\ \vdots & & \vdots & & \\ x_{n,1} & \cdots & x_{n,p_1} & & \end{pmatrix}$$

and these are called a sample with 2-step monotone missing data pattern.

The sample displayed in table 1 shows this pattern and it can be obtained by discarding the values of the last variable in the original one given in Little and Rubin(1987).

<Table 1> Sample with 2-step monotone missing data pattern

No. of observations	x_1	x_2	x_3	x_4
1	78.5	6.0	7.0	26.0
2	74.3	15.0	1.0	29.0
3	104.3	8.0	11.0	56.0
4	87.6	8.0	11.0	31.0
5	95.9	6.0	7.0	52.0
6	109.2	9.0	11.0	51.0
7	102.7	17.0	3.0	71.0
8	72.5	22.0	1.0	31.0
9	93.1	18.0	2.0	54.0
10	115.9	4.0		
11	83.8	23.0		
12	113.8	9.0		
13	109.4	8.0		

3. Parameter Estimation by Maximum Likelihood Method

The maximum likelihood estimates of the parameters can be obtained by finding those values that maximize the function given by

$$L(\mu_0, \mu_1, \Sigma_{00}, \Sigma_{11}) = \Pi_{i=1}^{n_1} f(x_{i,0}; \mu_0, \Sigma_{00}) \, \Pi_{i=n_1+1}^{n} f(x_{i,1}; \mu_1, \Sigma_{11}), \tag{3.1}$$

where $f(\cdot)$ is a multivariate normal density function. Due to the difficulty of manipulating the function given in this representation, we consider a more convenient approach than (3.1).

Let $x^{(1)} = (x_1, \cdots, x_{p_1})^t$ denote a subvector of x_0 containing the first p_1 variables and $x^{(2)} = (x_{p_1+1}, \cdots, x_p)^t$ denote a subvector of x_0 containing the remaining $p_2 (= p - p_1)$ variables. Assume that $x^{(i)} \sim N_{p_i}(\mu^{(i)}, \Sigma_{ii})$, $i = 1, 2$, where $\mu^{(i)}$ and Σ_{ii} are the mean vector and covariance matrix of $x^{(i)}$, and that two subvectors have the covariance matrix Σ_{12}. Denote the conditional likelihood function of $x_i^{(2)}$'s given $x_i^{(1)}$'s and the likelihood function of $x_i^{(1)}$'s by $L_1(\mu^* + Bx^{(1)}, \Sigma_{22.1}) = \Pi_{i=1}^{n_1} f(x_i^{(2)} \mid \mu^* + Bx_i^{(1)}, \Sigma_{22.1})$ and $L_2(\mu^{(1)}, \Sigma_{11}) = \Pi_{i=1}^{n} f(x_i^{(1)}; \mu^{(1)}, \Sigma_{11})$, respectively, where $x_i^{(1)} = (x_{i1}, \cdots, x_{i,p_1})^t$, $i = 1, \cdots, n$, is an observation on $x^{(1)}$ and $x_i^{(2)} = (x_{i,p_1+1}, \cdots, x_{i,p})^t$, $i = 1, \cdots, n_1$, is an observation on $x^{(2)}$, $\mu^* = \mu^{(2)} - B\mu^{(1)}$, $B = \Sigma_{21}\Sigma_{11}^{-1}$ and $\Sigma_{22.1} = \Sigma_{22} - B\Sigma_{11}B^t$.

Then, the overall likelihood function can be written by

$$L(\mu^{(1)}, \mu^{(2)}, \Sigma_{22.1}, \Sigma_{11}) = L_1(\mu^* + Bx^{(1)}, \Sigma_{22.1})\, L_2(\mu^{(1)}, \Sigma_{11}). \tag{3.2}$$

The maximum likelihood estimate of $\mu^{(1)}$ for fixed Σ_{11} is obtained by differentiating the second term of the right hand side of (3.2) with respect to $\mu^{(1)}$ and equating it to the null vector. This gives the maximum likelihood estimate of $\mu^{(1)}$ by $\hat{\mu}^{(1)} = \bar{x}^{(1)} = \sum_{i=1}^{n} x_i^{(1)}/n$. To get the estimate of Σ_{11}, the concentrated likelihood function given by

$$L_2(\bar{x}^{(1)}, \Sigma_{11}) = (2\pi)^{-np_1/2} |\Sigma_{11}|^{-n/2} \exp\left(-\frac{1}{2}\Sigma_{11}^{-1}\, tr\, W_{11}^*\right) \tag{3.3}$$

is maximized with respect to Σ_{11}, where $W_{11}^* = \sum_{i=1}^{n}\left(x_i^{(1)} - \bar{x}^{(1)}\right)\left(x_i^{(1)} - \bar{x}^{(1)}\right)^t$. When $\Sigma_{11}^{-1} W_{11}^* = nI$, the function (3.3) is maximized and thus the estimate of Σ_{11} is given by $\hat{\Sigma}_{11} = S_{11}^* = \sum_{i=1}^{n}\left(x_i^{(1)} - \bar{x}^{(1)}\right)\left(x_i^{(1)} - \bar{x}^{(1)}\right)^t/n$.

In a similar manner, the estimates of μ^*, B and $\Sigma_{22.1}$ can be obtained by maximizing the first term of the right hand side of (3.2),

$$L_1(\mu^* + Bx^{(1)}, \Sigma_{22.1}) = (2\pi)^{-n_1p_1/2} |\Sigma_{22.1}|^{-n_1/2} \exp\left(-\frac{1}{2}\sum_{i=1}^{n_1} d_i^t \Sigma_{22.1}^{-1} d_i\right), \tag{3.4}$$

where $d_i = (x_i^{(2)} - \mu^* - Bx_i^{(1)})^t$. Firstly, for fixed both B and $\Sigma_{22.1}$, μ^* maximizing the function (3.4) is obtained by $\hat{\mu}^* = \bar{x}_{n_1}^{(2)} - B\bar{x}_{n_1}^{(1)}$, where $\bar{x}_{n_1}^{(1)} = \sum_{i=1}^{n1} x_i^{(1)}/n_1$ and $\bar{x}_{n_1}^{(2)} = \sum_{i=1}^{n1} x_i^{(2)}/n_1$. Secondly, to get the estimate of B for fixed $\Sigma_{22.1}$, we differentiate the following concentrated log-likelihood function with respect to B,

$$l(\hat{\mu}^* + Bx^{(1)}, \Sigma_{22.1}) = -\frac{n_1p_1}{2}\log(2\pi) - \frac{n_1}{2}\log|\Sigma_{22.1}| - \frac{1}{2}\{tr(M_1 - M_2 - M_3 + M_4)\}, \tag{3.5}$$

where $M_1 = \Sigma_{22.1}^{-1} W_{22}$, $M_2 = \Sigma_{22.1}^{-1} B W_{12}$, $M_3 = B^t \Sigma_{22.1}^{-1} W_{21}$, $M_4 = B^t \Sigma_{22.1}^{-1} B W_{11}$ and W_{ij}, $i,j = 1,2$, is the sum of squares and product matrix formed from $x_i^{(1)}$ and $x_i^{(2)}$, $i = 1,\cdots,n_1$, given by $W_{11}=\sum_{i=1}^{n_1}\left(x_i^{(1)}-\bar{x}_{n_1}^{(1)}\right)\left(x_i^{(1)}-\bar{x}_{n_1}^{(1)}\right)^t$, $W_{12}=\sum_{i=1}^{n_1}\left(x_i^{(1)}-\bar{x}_{n_1}^{(1)}\right)\left(x_i^{(2)}-\bar{x}_{n_1}^{(2)}\right)^t$, $W_{21}=W_{12}^t$ and $W_{22}=\sum_{i=1}^{n_1}\left(x_i^{(2)}-\bar{x}_{n_1}^{(2)}\right)\left(x_i^{(2)}-\bar{x}_{n_1}^{(2)}\right)^t$, respectively. By solving the likelihood equation $\partial l/\partial B=\Sigma_{22.1}^{-1}(W_{21}-BW_{11})=0$, we obtain the estimate of B by $\hat{B}=W_{21}W_{11}^{-1}$. Finally, the estimate of $\Sigma_{22.1}$ maximizing $L_1(\hat{\mu}^*+\hat{B}x^{(1)},\Sigma_{22.1})$ with respect to $\Sigma_{22.1}$ is given by $\hat{\Sigma}_{22.1}=S_{22.1}=(W_{22}-\hat{B}W_{11}\hat{B}^t)/n_1$.

The maximum likelihood estimates of other parameters can be obtained by solving the relations $\mu^*=\mu^{(2)}-B\mu^{(1)}$, $B=\Sigma_{21}\Sigma_{11}^{-1}$ and $\Sigma_{22.1}=\Sigma_{22}-B\Sigma_{11}B^t$ and applying Corollary 3.2.1 of Anderson(1984).

4. Likelihood Ratio Criterion for Testing Sphericity

To derive the likelihood ratio test criterion for the sphericity hypothesis $H_0:\Sigma=\sigma^2I$, the parameters must be estimated under the null hypothesis. Since the covariance matrix Σ can be decomposed into $\Sigma_{11}=\sigma^2I_{p_1}$, $\Sigma_{22}=\sigma^2I_{p_2}$ and $\Sigma_{12}=\Sigma_{21}^t=0$ under the null hypothesis, the parameters estimated by the maximum likelihood method are $\mu^{(1)}$, $\mu^{(2)}$ and σ^2.

The likelihood function based on n observations $x_i^{(1)}$'s and n_1 observations $x_i^{(2)}$'s can be written

$$L_{H_0}(\mu^{(1)},\mu^{(2)},\sigma^2)=(2\pi)^{-\frac{np_1+n_1p_2}{2}}(\sigma^2)^{-\frac{np_1+n_1p_2}{2}}\exp\left(-\frac{q_1+q_2}{2\sigma^2}\right), \tag{4.1}$$

where $q_1=\sum_{i=1}^{n}(x_i^{(1)}-\mu^{(1)})^t(x_i^{(1)}-\mu^{(1)})$ and $q_2=\sum_{i=1}^{n_1}(x_i^{(2)}-\mu^{(2)})^t(x_i^{(2)}-\mu^{(2)})$. The estimates of the parameters are obtained by differentiating the log-likelihood function $l_{H_0}(\mu^{(1)},\mu^{(2)},\sigma^2)$ with respect to $\mu^{(1)}$, $\mu^{(2)}$ and σ^2 and solving the likelihood equations $\partial l_{H_0}(\mu^{(1)},\mu^{(2)},\sigma^2)/\partial\mu^{(1)}=0$, $\partial l_{H_0}(\mu^{(1)},\mu^{(2)},\sigma^2)/\partial\mu^{(2)}=0$ and $\partial l_{H_0}(\mu^{(1)},\mu^{(2)},\sigma^2)/\partial\sigma^2=0$. The maximum likelihood estimate of each parameter is given by

$$\hat{\mu}^{(1)}=\bar{x}^{(1)}=\sum_{i=1}^{n}x_i^{(1)}/n,\ \hat{\mu}^{(2)}=\bar{x}_{n_1}^{(2)}=\sum_{i=1}^{n_1}x_i^{(2)}/n_1,\ \hat{\sigma^2}=s^2=\frac{w_{11}^0+w_{22}^0}{np_1+n_1p_2}, \tag{4.2}$$

where $w_{11}^0 = \sum_{i=1}^{n} \left(x_i^{(1)} - \bar{x}^{(1)}\right)^t \left(x_i^{(1)} - \bar{x}^{(1)}\right)$ and $w_{22}^0 = \sum_{i=1}^{n_1} \left(x_i^{(2)} - \bar{x}_{n_1}^{(2)}\right)^t \left(x_i^{(2)} - \bar{x}_{n_1}^{(2)}\right)$. By substituting the estimates of (4.2) into (4.1), we have the maximum value of the likelihood function under the null hypothesis,

$$L_{H_0}\left(\bar{x}^{(1)}, \bar{x}_{n_1}^{(2)}, s^2\right) = (2\pi)^{-\frac{np_1 + n_1 p_2}{2}} \left(\frac{w_{11}^0 + w_{22}^0}{np_1 + n_1 p_2}\right)^{-\frac{np_1 + n_1 p_2}{2}} \exp\left(-\frac{np_1 + n_1 p_2}{2}\right). \tag{4.3}$$

In the case of the alternative hypothesis, the maximum value of the likelihood function after substituting the estimates of the parameters derived in section 3 is given by

$$L_{H_1}\left(\bar{x}^{(1)}, \bar{x}_{n_1}^{(2)}, S_{11}^*, S_{22.1}\right) = (2\pi)^{-\frac{np_1 + n_1 p_2}{2}} |S_{11}^*|^{-\frac{n}{2}} |S_{22.1}|^{-\frac{n_1}{2}} \exp\left(-\frac{np_1 + n_1 p_2}{2}\right). \tag{4.4}$$

Taking the ratio of $L_{H_0}\left(\bar{x}^{(1)}, \bar{x}_{n_1}^{(2)}, s^2\right)$ to $L_{H_1}\left(\bar{x}^{(1)}, \bar{x}_{n_1}^{(2)}, S_{11}^*, S_{22.1}\right)$, we get the likelihood ratio test statistic

$$\lambda = \frac{|S_{11}^*|^{\frac{n}{2}} |S_{22.1}|^{\frac{n_1}{2}}}{\left(\frac{w_{11}^0 + w_{22}^0}{np_1 + n_1 p_2}\right)^{\frac{np_1 + n_1 p_2}{2}}} = \frac{|S_{11}^*|^{\frac{n}{2}} |S_{22.1}|^{\frac{n_1}{2}}}{\left(\frac{n\,tr S_{11}^* + n_1 tr S_{22.1}}{np_1 + n_1 p_2}\right)^{\frac{np_1 + n_1 p_2}{2}}}. \tag{4.5}$$

The null hypothesis is rejected in favor of the alternative when the test statistic λ is less than or equal to a constant $c(\alpha)$, where $c(\alpha)$ is a critical value depending on the significance level of the test. However, it is difficult to derive the exact null distribution of the test statistic and thus we may use the asymptotic result that $\lambda^* = -2\log\lambda$ follows the chi-squared distribution with degrees of freedom $(p+2)(p-1)/2$ as overall sample size increases. Thus, the null hypothesis is rejected at the significance level α if $\lambda^* \geq c^*(\alpha)$, where $c^*(\alpha)$ is the 100α percentile of the chi-squared distribution.

5. Numerical Example

The sample given in table 1 is used for a numerical illustration to implement the testing procedure proposed in the previous section. In the sample, $x^{(1)} = (x_1, x_2)^t$, $x^{(2)} = (x_3, x_4)^t$ and $x_0 = (x_1, x_2, x_3, x_4)^t$. Also, $n_1 = 9$, $n_2 = 4$, $n = 13$ and $p_1 = p_2 = 2$. The numerical values of all the estimates were calculated by SAS/IML software.

The estimate of $x^{(1)}$ calculated from all observations is $\bar{x}^{(1)} = (95.462, 11.769)^t$. From 9 observations on x_0, two sample mean vectors, $x^{(1)}_{n_1}$ and $x^{(2)}_{n_1}$, are calculated as $x^{(1)}_{n_1} = (90.900, 12.111)^t$ and $x^{(2)}_{n_1} = (6.000, 44.556)^t$. Also, each submatrix of W is calculated as

$$W_{11} = \begin{pmatrix} 1462.300 & -212.500 \\ -212.500 & 282.890 \end{pmatrix}, \; W_{21} = \begin{pmatrix} 265.400 & -171.000 \\ 1423.800 & 62.444 \end{pmatrix}, \; W_{22} = \begin{pmatrix} 152.000 & 39.000 \\ 39.000 & 1972.200 \end{pmatrix}.$$

Using these values, we obtain the estimates of B and $\Sigma_{22.1}$,

$$\hat{B} = \begin{pmatrix} 0.105 & -0.526 \\ 1.129 & 1.069 \end{pmatrix}, \; S_{22.1} = \begin{pmatrix} 3.804 & -8.652 \\ -8.652 & 32.892 \end{pmatrix}.$$

Based on S^*_{11} calculated from all observations,

$$S^*_{11} = \begin{pmatrix} 210.300 & -47.660 \\ -47.660 & 37.870 \end{pmatrix},$$

we have

$$|S_{11}^*| = 5692.585, \; |S_{22.1}| = 50.264,$$
$$tr S^*_{11} = 248.170, \; tr S_{22.1} = 36.696.$$

The value of the test statistic is $\lambda = 1.255 \times 10^{-10}$ and thus we get $\lambda^* = 45.597$. Since its p-value is $P(\lambda^* \geq 45.597) = P(\chi^2_9 \geq 45.597) = 7.15 \times 10^{-7}$, the null hypothesis is rejected at the significance level 5%.

6. Conclusion

In this paper, we introduced the testing problem for sphericity when there is a multivariate normal sample with 2-step monotone missing data pattern. The likelihood ratio test criterion derived is based on the asymptotic null distribution of λ^*. The form of the test statistic appears to be somewhat different from the usual one constructed from complete observations. This leads to the difficulty of deriving the exact null distribution of the test statistic by an approach commonly used in multivariate statistical theory. Though the exact null distribution could not presented in this paper for that reason, this problem will be throughly investigated in the future.

Finally, in the situation where a given sample takes the form of 2-step monotone missing data pattern, the test presented is expected to be useful in applications where sphericity plays a key role in.

References

[1] Amey, A.K.A. and Gupta, A.K.(1992). Testing sphericity under a mixture model, *Australian Journal of Statistics*, Vol. 34, 451-460.

[2] Anderson, T.W.(1957). Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association*, Vol. 52, 200-203.

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

[4] Anderson, T.W. and Olkin, I.(1985). Maximum-likelihood estimation of the parameters of a multivariate normal distribution, *Linear Algebra and its Applications*, Vol. 70, 147-171.

[5] Bhargava, R.P.(1975). Some one-sample hypothesis testing problems when there is a monotone sample from a multivariate normal population, *Annals of the Institute of Statistical Mathematics*, Vol. 27, 327-339.

[6] Eaton, M.L. and Kariya, T.(1983). Multivariate tests with incomplete data, *Annals of Statistics*, Vol. 11, 654-665.

[7] Gupta, A.K.(1977). On the distribution of sphericity test criterion in the multivariate Gaussian distribution, *Australian Journal of Statistics*, Vol. 19, 202-205.

[8] Jinadasa, K.G. and Tracy, D.S.(1992). Maximum likelihood estimation for multivariate normal distribution with monotone sample, *Communications in Statistics-Theory and Methods*, Vol. 21, 41-50.

[9] John, S.(1972). The distribution of a statistic used for testing sphericity of normal distributions, *Biometrika*, Vol. 59, 169-174.

[10] Little, R.J.A. and Rubin, D.B.(1987). *Statistical Analysis with Missing Data*, Wiley, New York.

[11] Lord, F.M.(1955). Estimation of parameters from incomplete data, *Journal of the American Statistical Association*, Vol. 50, 870-876.

[12] Mardia, K.V. and Jupp, P.E.(1999). *Directional Statistics*, Wiley, New York.

[13] Mauchly, J.W.(1940). Significance test for sphericity of a normal n-variate distribution, *Annals of Mathematical Statistics*, Vol. 11, 204-209.

[14] Nagar, D.K., Jain, S.K. and Gupta, A.K.(1991). Distribution of LRC for testing sphericity structure of a covariance matrix in multivariate normal distribution, *Metron*, Vol. 49,

435-457.

[15] Provost, S.B.(1990). Estimators for the parameters of a multivariate normal random vector with incomplete data on two subvectors and test of independence, *Computational Statistics and Data Analysis*, Vol. 9, 37-46.

[16] Sugira, N.(1972). Locally best invariant test for sphericity and the limiting distributions, *Annals of Mathematical Statistics*, Vol. 43, 1312-1316.

[Received February 2005, Accepted June 2005]