Influence Measures for the Likelihood Ratio Test on Independence of Two Random Vectors

Kang-Mo Jung[1]

Abstract

We compare methods for detecting influential observations that have a large influence on the likelihood ratio test statistic that the two sets of variables are uncorrelated with one another. For this purpose we derive results of the deletion diagnostic, the influence function, the standardized influence matrix and the local influence. An illustrative example is given.

Keywords: Comparing covariance matrices, deletion, influence function, likelihood ratio test, local influence, standardized influence matrix.

1. Introduction

The detection of outliers or influential observations has a long history. However, many diagnostic measures have been proposed for influence analyses in the context of estimation. A few works that treat detection of influential observations for test statistics in multivariate analysis are found. Among others, Jung (2001) investigated the influence of observations on the likelihood ratio test (LRT) statistics in the canonical correlation analysis using the local influence method introduced by Cook (1986). Influence analysis in testing problems is very important because in extreme situations, few observations can dominate our conclusion about the hypothesis as can be seen in Section 3.

Assume that the random vector $\mathbf{z} = (\mathbf{x}^T, \mathbf{y}^T)^T$ has the covariance matrix $\mathbf{\Sigma}$, where $\mathbf{\Sigma}$ is partitioned such that

$$\mathbf{\Sigma} = \begin{pmatrix} \mathbf{\Sigma}_{11} & \mathbf{\Sigma}_{12} \\ \mathbf{\Sigma}_{21} & \mathbf{\Sigma}_{22} \end{pmatrix}.$$

And that $\mathbf{x}$ and $\mathbf{y}$ are p and q dimensional random vectors, respectively.

Consider the hypothesis

$$H_0 : \mathbf{\Sigma}_{12} = \mathbf{0}, \tag{1}$$

which means the two sets of variables are uncorrelated with one another. Under the normality, the LRT statistic for testing H_0 is given by

$$T = -(n - (p + q + 3)/2) \log \frac{|\mathbf{S}|}{|\mathbf{S}_{11}||\mathbf{S}_{22}|}, \tag{2}$$

where $\mathbf{S}, \mathbf{S}_{11}, \mathbf{S}_{22}$ is the maximum likelihood estimators (MLE) of $\mathbf{\Sigma}, \mathbf{\Sigma}_{11}, \mathbf{\Sigma}_{22}$, respectively, and n is the number of observations. Then the test statistic T

[1] Department of Informatics & Statistics, Kunsan National University, 68 Miryong-Dong, Kunsan, Chollapuk-Do, 573–701, Korea

is approximately distributed as a chi-squared distribution with pq degrees of freedom (Mardia, et al., 1979, p. 288 and Rencher, 1995, pp. 285 -287).

It is well known that the sample covariance matrix is very sensitive to outliers (Critchley, 1985), and so the LRT statistic T. To investigate the influence of observations on the test statistic (??), Jung (2001) considered the local influence method using the fact that the test statistic can be rewritten by the squared canonical correlation coefficients. In this work we considered the deletion approach, the influence function and the standardized influence matrix.

In Section 2 we will derive the deletion diagnostic, the influence function and the standardized influence matrix of T. In Section 3 a numerical example will be given for illustration.

2. Influence Measures

The random sample $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ is drawn from $(p+q)-$variate normal distribution $\mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. Assume that $\mathbf{z}_u$ is decomposed as in Section 1, that is, $\mathbf{z}_u = (\mathbf{x}_u^T, \mathbf{y}_u^T)^T$. Then MLE of $\boldsymbol{\Sigma}$ becomes $\mathbf{S} = (1/n)\sum_{i=1}^n (\mathbf{z}_i - \bar{\mathbf{z}})(\mathbf{z}_i - \bar{\mathbf{z}})^T$. Also MLE of $\boldsymbol{\Sigma}_{11}$ and $\boldsymbol{\Sigma}_{22}$ are similarly obtained.

2.1. Deletion Diagnostic

We will derive the deletion diagnostic for the test statistic T. Let $T_{(u)}$ be the test statistic with the deletion of uth observation $\mathbf{z}_u$. Hereafter we denote by the subscript (u) the estimator or statistic based on the reduced data set without observation $\mathbf{z}_u$. Since

$$\mathbf{S}_{(u)} = \frac{n}{n-1}\left[\mathbf{S} - \frac{1}{n-1}(\mathbf{z}_u - \bar{\mathbf{z}})(\mathbf{z}_u - \bar{\mathbf{z}})^T\right],$$

$$|\mathbf{S}_{(u)}| = \left(\frac{n}{n-1}\right)^{p+q} |\mathbf{S}| \left(1 - \frac{D_z}{n-1}\right),$$

where $D_z = (\mathbf{z}_u - \bar{\mathbf{z}})^T \mathbf{S}^{-1} (\mathbf{z}_u - \bar{\mathbf{z}})$. Similarly we obtain $|\mathbf{S}_{11(u)}|$ and $|\mathbf{S}_{22(u)}|$, where D_x and D_y are similarly defined as D_z. Then we have

$$T_{(u)} = \left(1 + \frac{1}{c}\right) T + c^* \left\{\log\left(1 - \frac{D_z}{n-1}\right) - \log\left(1 - \frac{D_x}{n-1}\right) - \log\left(1 - \frac{D_y}{n-1}\right)\right\}, \tag{3}$$

where $c^* = c - 1$ and $c = -(n - (p+q+3)/2)$.

¿From (??), we obtain directly the deletion diagnostic $T_{(u)} - T$ for the LRT statistic.

2.2. Influence Function

The influence function (Hampel, 1974) for a parameter at a point measures the effect of an infinitesimal contamination at that point on the estimation of

the parameter. Hence the influence function can serve as a diagnostic method of detecting influential observations in performing a test of hypothesis.

Using the approach of Jung and Kim (1999), we can construct the statistical functional $T(F) = c \log \tau$, where $\tau = |\mathbf{\Sigma}|/(|\mathbf{\Sigma}_{11}||\mathbf{\Sigma}_{22}|)$. Then we obtain the empirical influence function of T as

$$\begin{aligned} EIF(\mathbf{z}_u, T) &= \frac{c}{\tau} EIF(\mathbf{z}_u, \tau) \\ &= D_z - D_x - D_y \end{aligned} \tag{4}$$

from Radhakrishnan and Kshirsagar (1981).

Equation (??) can be rewritten as

$$EIF(\mathbf{z}_u, T) = (\mathbf{z}_u - \bar{\mathbf{z}})^T (\mathbf{S}^{-1} - \mathbf{S}_0^{-1})(\mathbf{z}_u - \bar{\mathbf{z}}), \tag{5}$$

where $\mathbf{S}_0$ is the MLE of covariance matrix under (??) H_0.

2.3. Standardized Influence Matrix

Lu et al. (1997) extended the influence function to the standardized influence matrix (SIM) of an estimator. For the statistical functional T they defined the SIM as

$$SIM(\mathbf{Z}, T, F) = \frac{1}{n} IF(\mathbf{Z}, T, F)^T \mathbf{V}^{-1} IF(\mathbf{Z}, T, F),$$

where $\mathbf{Z}$ is the data matrix and $\mathbf{V}$ is the covariance matrix of $IF(\mathbf{Z}, T, F)$. Also they suggested a diagnostic method by plotting the eigenvector corresponding to the largest eigenvalue of SIM.

In this work, the dimension of statistical functional T is 1 and so the sample version of $\mathbf{V}$ is scalar. We consider $EIF(\mathbf{Z}, T) EIF(\mathbf{Z}, T)^T$ as the sample version $\widehat{SIM}$ of SIM. Then the eigenvector of $\widehat{SIM}$ becomes the standardized vector of $EIF(\mathbf{Z}, T)$. It indicates that the influence information of SIM and the influence function give the same influence information.

3. Numerical Example

Diagnostic measures described in Section 2 are applied to the head-length data (Mardia, et al., 1979, p. 121, Table 5.1.1) previously analyzed by Jung (2001). For this data set, $n = 25, p = 2, q = 2$. The LRT statistic based on the full data set is $T = 20.96$, and therefore we conclude that the null hypothesis is strongly rejected from the $p-$value 0.003.

We obtain information about influential observations for the LRT statistic T using the deletion diagnostic (??), the influence function (??), the standardized influence matrix and the local influence method. The result of local influence method is obtained from Jung (2001). Since the results of the influence function and SIM are same, we omit those of SIM. The deletion diagnostic and influence function are the standardized version of $T - T_{(u)}$ and EIF, respectively. The results are plotted in Figure 1.

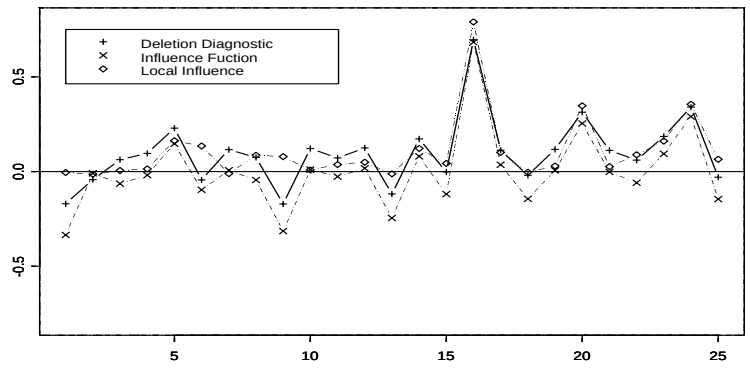

Figure 1: The index plots for the LRT statistic T using influence measures

¿From Figure 1 we may conclude that observation 16 is most influential from all influence measures. Observations 20 and 24 are candidate for influential observations. The influence of observations 16, 20 and 24 are confirmed by $p-$value 0.068 of the LRT statistic with the remaining data set discarding those observations.

References

1. Cook, R.D. (1986). Assessment of local influence (with discussions), *J. Roy. Statist. Soc. B*, **48**, 133–169.
2. Critchley, F. (1985). Influence in principal component analysis, *Biometrika*, **72**, 627–636.
3. Hampel, F. R. (1974). The influence curve and its role in robust estimation, *J. Amer. Statist. Assoc.*, **69**, 383–393.
4. Jung, K.-M. (2001). Influence analysis on a test statistic in canonical correlation coefficients, *The Korean Commun. in Statist.*, **8**, 347–355.
5. Jung, K.-M. and Kim, M. G. (1999). Influence analysis of the likelihood ratio test in multivariate Behrens-Fisher problem, *The Korean Commun. in Statist.*, **6**, 939–946.
6. Mardia, K. V., Kent, J. T. and Bibby, J. M. (1979). *Multivariate Analysis*, Academic Press, New York.
7. Lu,, J., Ko., D. and Chang, T. (1997). The standardized influence matrix and its applications, *J. Amer. Statist. Assoc.*, **92**, 1572–1580.
8. Radhakrishnan, R. and Kshirsagar, A. M. (1981). Influence functions for certain parameters in multivariate analysis, *Commun. Statist.-Theory and Methods*, **10**(6), 515–529.
9. Rencher, A.C. (1995). *Methods of Multivariate Analysis*, Wiley, New York.