J. Appl. Math. & Informatics Vol. **27***(2009), No. 5 - 6, pp. 1429 - 1433*
Website: http://www.kcam.biz

MULTIVARIATE JOINT NORMAL LIKELIHOOD DISTANCE

MYUNG GEUN KIM

ABSTRACT. The likelihood distance for the joint distribution of two multivariate normal distributions with common covariance matrix is explicitly derived. It is useful for identifying outliers which do not follow the joint multivariate normal distribution with common covariance matrix. The likelihood distance derived here is a good ground for the use of a generalized Wilks statistic in influence analysis of two multivariate normal data.

AMS Mathematics Subject Classification : 62J20.
Key words and phrases : Influence, likelihood distance, multivariate normal distribution.

1. Introduction

Two multivariate normal distributions with common covariance matrix are often adopted in multivariate data analysis. One of the two typical cases is Hotelling's T^2 test for comparing two mean vectors. The other is the linear discriminant analysis. In this situation, no literature is available for detecting outliers which do not follow the joint multivariate normal distribution with common covariance matrix. Wilks statistic (Wilks[7]) suggested for a single population case is easily generalized to the situation of two multivariate normal distributions with common covariance matrix and therefore it can be a candidate for checking the outlyingness of observations from two multivariate normal distributions with common covariance matrix. However, no explicit exposition or ground for the use of Wilks statistic in this situation is available.

In this work the likelihood distance is considered for two multivariate normal distributions with common covariance matrix. It is explicitly derived. It is useful for investigating the influence of observations on the multivariate joint normal likelihood. It will be seen that the likelihood distance should be a good ground for the use of Wilks statistic in influence analysis of two multivariate normal data.

Received January 13, 2009. Accepted March 5, 2009.

2. Likelihood distance

Let $x_{i1},...,x_{i,n_i}$ be a random sample of size n_i from the multivariate normal distribution $N(\mu_i, \Sigma)$ $(i = 1, 2)$. For each distribution $N(\mu_i, \Sigma)$, the maximum likelihood estimators of μ_i and Σ based on each sample are denoted by the sample mean vector $\bar{x}_i$ and the sample covariance matrix S_i, respectively which are given by

$$\begin{aligned} \bar{x}_i &= \frac{1}{n_i}\sum_{j=1}^{n_i} x_{ij} \\ S_i &= \frac{1}{n_i}\sum_{j=1}^{n_i}(x_{ij}-\bar{x}_i)(x_{ij}-\bar{x}_i)^T. \end{aligned}$$

For the joint distribution of $N(\mu_1, \Sigma)$ and $N(\mu_2, \Sigma)$, the maximum likelihood estimator of Σ based on the two samples is denoted by the pooled sample covariance matrix S given by $S = (n_1S_1 + n_2S_2)/n$ where $n = n_1 + n_2$. Then the joint log-likelihood function of μ_1, μ_2 and Σ can be written as

$$l(\mu_1, \mu_2, \Sigma) = -\frac{n}{2}\{\log \mid 2\pi\Sigma \mid + \mathrm{tr}(\Sigma^{-1}S)\} - \sum_{i=1}^{2}\frac{n_i}{2}(\bar{x}_i-\mu_i)^T\Sigma^{-1}(\bar{x}_i-\mu_i).$$

We let

$$\begin{aligned} \bar{x}_{i(r)} &= \frac{1}{n_i-1}\sum_{j\neq r} x_{ij} \\ &= \frac{n_i}{n_i-1}\bar{x}_i - \frac{1}{n_i-1}x_{ir} \\ S_{i(r)} &= \frac{1}{n_i-1}\sum_{j\neq r}(x_{ij}-\bar{x}_{i(r)})(x_{ij}-\bar{x}_{i(r)})^T \\ &= \frac{n_i}{n_i-1}S_i - \frac{n_i}{(n_i-1)^2}(x_{ir}-\bar{x}_i)(x_{ir}-\bar{x}_i)^T. \end{aligned}$$

A little computation shows that the second equalities for $\bar{x}_{i(r)}$ and $S_{i(r)}$ hold.

2.1. The influence of an observation in the first sample

First, we will derive a formula which measures the influence of an observation x_{1r} in the first sample on the joint likelihood function. When the r-th observation x_{ir} is omitted from the first sample, the maximum likelihood estimators of μ_1, μ_2 and Σ based on the remaining sample of size $n-1$ becomes $\bar{x}_{1(r)}$, $\bar{x}_2$ and $S_{(1r)}$, respectively, where

$$S_{(1r)} = \frac{1}{n-1}[(n_1-1)S_{1(r)} + n_2S_2].$$

Then the likelihood distance is defined by

$$LD_{1r} = 2\Big[l(\bar{x}_1, \bar{x}_2, S) - l(\bar{x}_{1(r)}, \bar{x}_2, S_{(1r)})\Big] \quad (r = 1, ..., n_1)$$

which measures the influence of x_{1r} on the joint likelihood.

In order to get the likelihood distance LD_{1r}, first we compute $l(\bar{x}_1, \bar{x}_2, S)$ which results in

$$l(\bar{x}_1, \bar{x}_2, S) = -\frac{n}{2}\{p\log(2\pi) + \log \mid S \mid + p\}.$$

Next, we need some preliminary computations to get $l\Big(\bar{x}_{1(r)}, \bar{x}_2, S_{(1r)}\Big)$. An explicit expression for $S_{(1r)}$ is easily found as

$$S_{(1r)} = \frac{n}{n-1}S - \frac{n_1}{(n-1)(n_1-1)}(x_{1r} - \bar{x}_1)(x_{1r} - \bar{x}_1)^T.$$

Using (A.2.3n) in p.458 of Mardia et al.[6], it is easily to find

$$\mid S_{(1r)} \mid = \left(\frac{n}{n-1}\right)^p \mid S \mid \left[1 - \frac{n_1}{n(n_1-1)}D_{1r}^2\right]$$

where $D_{1r}^2 = (x_{1r} - \bar{x}_1)^T S^{-1}(x_{1r} - \bar{x}_1)$. Using (A.2.4f) in p.459 of Mardia et al.[6], we have

$$S_{(1r)}^{-1} = \frac{n-1}{n}\left[I + \frac{n_1 S^{-1}(x_{1r} - \bar{x}_1)(x_{1r} - \bar{x}_1)^T}{n(n_1-1) - n_1 D_{1r}^2}\right] S^{-1}$$

which yields

$$\begin{aligned} \mathrm{tr}(S_{(1r)}^{-1}S) &= \frac{n-1}{n}\left[p + \frac{n_1 D_{1r}^2}{n(n_1-1) - n_1 D_{1r}^2}\right] \\ (\bar{x}_1 - \bar{x}_{1(r)})^T S_{(1r)}^{-1}(\bar{x}_1 - \bar{x}_{1(r)}) &= \left(\frac{n-1}{n_1-1}\right)\frac{D_{1r}^2}{n(n_1-1) - n_1 D_{1r}^2}. \end{aligned}$$

Thus we get

$$\begin{aligned} l(\bar{x}_{1(r)}, \bar{x}_2, S_{(1r)}) &= -\frac{n}{2}\{p\log(2\pi) + p\log(\frac{n}{n-1}) + \log\left(1 - \frac{n_1}{n(n_1-1)}D_{1r}^2\right) \\ &\quad + \log \mid S \mid + \frac{n-1}{n}\left(p + \frac{n_1 D_{1r}^2}{n(n_1-1) - n_1 D_{1r}^2}\right)\} \\ &\quad - \frac{n_1}{2}\left(\frac{n-1}{n_1-1}\right)\frac{D_{1r}^2}{n(n_1-1) - n_1 D_{1r}^2}. \end{aligned}$$

Hence the likelihood distance becomes

$$\begin{aligned} LD_{1r} &= np\log(\frac{n}{n-1}) - p + n\log\left(1 - \frac{n_1}{n(n_1-1)}D_{1r}^2\right) \\ &\quad + \left(\frac{n_1^2(n-1)}{n_1-1}\right)\frac{D_{1r}^2}{n(n_1-1) - n_1 D_{1r}^2}. \end{aligned} \tag{1}$$

The LD_{1r} is invariant under affine transformation of x_{ij}. Since it is easily seen that LD_{1r} is a strictly increasing function of D_{1r}^2, an investigation of the

influence of x_{1r} on the joint likelihood function is equivalent to uncovering the outlyingness of x_{1r} based on D_{1r}^2.

2.2. The influence of an observation in the second sample

In order to investigate the influence of x_{2s} in the second sample on the joint likelihood, similarly to LD_{1r} in (1) we can define the likelihood distance by

$$LD_{2s} = 2\Big[l(\bar{x}_1, \bar{x}_2, S) - l(\bar{x}_1, \bar{x}_{2(s)}, S_{(2s)})\Big] \quad (s = 1, ..., n_2)$$

where

$$\begin{aligned} S_{(2s)} &= \frac{1}{n-1}[n_1 S_1 + (n_2 - 1)S_{2(s)}] \\ &= \frac{n}{n-1}S - \frac{n_2}{(n-1)(n_2-1)}(x_{2s} - \bar{x}_2)(x_{2s} - \bar{x}_2)^T. \end{aligned}$$

Symmetric role of notations enables us to get LD_{2s} by directly substituting n_2 and D_{2s}^2 for n_1 and D_{1r}^2 in (1), where $D_{2s}^2 = (x_{2s} - \bar{x}_2)^T S^{-1}(x_{2s} - \bar{x}_2)$. First, we get

$$\begin{aligned} l(\bar{x}_1, \bar{x}_{2(s)}, S_{(2s)}) &= -\frac{n}{2}\Big\{p\log(2\pi) + p\log(\frac{n}{n-1}) + \log\left(1 - \frac{n_2}{n(n_2-1)}D_{2s}^2\right) \\ &\quad + \log \mid S \mid + \frac{n-1}{n}\left(p + \frac{n_2 D_{2s}^2}{n(n_2-1) - n_2 D_{2s}^2}\right)\Big\} \\ &\quad - \frac{n_2}{2}\left(\frac{n-1}{n_2-1}\right)\frac{D_{2s}^2}{n(n_2-1) - n_2 D_{2s}^2}. \end{aligned}$$

Then the likelihood distance becomes

$$\begin{aligned} LD_{2s} &= np\log(\frac{n}{n-1}) - p + n\log\left(1 - \frac{n_2}{n(n_2-1)}D_{2s}^2\right) \\ &\quad + \left(\frac{n_2^2(n-1)}{n_2-1}\right)\frac{D_{2s}^2}{n(n_2-1) - n_2 D_{2s}^2}. \end{aligned} \tag{2}$$

Since LD_{2s} is also a strictly increasing function of D_{2s}^2, an investigation of the influence of x_{2s} on the joint likelihood function is equivalent to uncovering the outlyingness of x_{2s} based on D_{2s}^2.

3. Discussions

The results show that an investigation of the influence of observations based on the likelihood distances LD_{1r} or LD_{2s} is equivalent to measuring the influence based on D_{1r}^2 or D_{2s}^2, respectively, each of which is a generalizations of Wilks statistic to the joint distribution with common covariance matrix.

Wilks statistic (Wilks[7]) was suggested for a single population case and it appears in wide areas, for example in Kim[4]. Its generalization can be used

for checking the outlyingness of observations from two multivariate normal distributions with common covariance matrix. The results in Section 2 is a good ground for the use of Wilks statistic in this situation.

The influence measures D_{1r}^2 or D_{2s}^2 are found in many situations. Among others, Kim[3] derived influence functions for comparing covariance matrices where the influence measures D_{ij}^2 are used in investigating the influence in comparing covariance matrices. A generalization of local influence to the case of two multivariate normal distributions with common covariance matrix also contains D_{ij}^2 as components. Influence analysis of Hotelling's test statistic includes D_{ij}^2 as components (Kim[5]). The influence measures D_{ij}^2 appear in local influence analysis of linear discriminant problem (Jung et al.[1]), and also in influence analysis of selecting discriminant variables (Jung and Kim[2]).

References

1. K.-M. Jung, M. G. Kim and B. C. Kim, *Second order local influence in linear discriminant analysis*, J. Jpn. Soc. Comp. Statist., **10** (1997), 1–11
2. K.-M. Jung and M. G. Kim, *Influence analysis in selecting discriminant variables*, J. Korean Statist. Soc., **30** (2001), 499–509
3. M. G. Kim, *The influence in comparing covariance matrices*, Commun. Statist.- Theory Meth., **24** (1995), 1431–1441.
4. M. G. Kim, *Case influence on multiple correlation coefficient*, J. Appl. Math. & Computing, **19** (2005), 521–525
5. M. G. Kim, *Influence on Hotelling's test statistic*, J. Korean Data Anal. Soc., **9** (2007), 545–551.
6. K. V. Mardia, J. T. Kent and J. M. Bibby, *Multivariate Analysis*, Academic Press, 1979
7. S. S. Wilks, *Multivariate statistical outliers*, Sankhyā, **25** (1963), 407–426

M.G. Kim received his Ph.D from Ohio State University. He is now a professor of Mathematics Education Department at Seowon University. His research interest centers on diagnostics in multivariate analysis and linear model.

Department of Mathematics Education, Seowon University, 231 Mochung-Dong, Heungduk-Gu, Cheongju, Chung-Buk, 361-742, Korea
e-mail : mgkim@seowon.ac.kr