응용통계연구 제 7권 2호
1994년 9월 pp. 213-224

Influence in Testing the Equality of Two Covariance Matrices

Myung Geun Kim[1]

Abstract

A diagnostic method useful for detecting outliers in testing the equality of two covariance matrices is developed using the influence curve approach. This method is easily generalized to more than two covariance matrices. A sample version for the influence measure of detecting outliers is considered based on the empirical distribution functions. The sample version includes as its component terms the well-known test statistic for detecting one outlier at a time introduced by Wilks and its generalization to the two-group case.

1. Introduction

A first step in comparing two covariance matrices is usually to check whether they are equal or not. Several procedures are available for testing the equality (see Seber, 1984, Chapter 3). However, no direct method of identifying influential observations in testing the equality of two covariance matrices is available. In such a case it will be interesting to develop a measure of detecting influential observations.

The influence curve introduced by Hampel (1974) has been used for detecting outliers in some areas, for example, principal component analysis (Critchley, 1985), discriminant analysis (Campbell, 1978) and regression diagnostics (Cook and Weisberg, 1982). The influence curve is usually defined for a parameter which can be expressed as a functional of the underlying distribution function, and it measures the effect of a point (an observation) on the parameter of interest.

One of test statistics for the equality of covariance matrices is the likelihood ratio statistic. This test statistic is a function of sample covariance matrices, and a sample covariance matrix is a functional evaluated at the corresponding empirical distribution function. Thus we can define an appropriate functional so that its influence curve can serve as a diagnostic method of identifying influential observations in testing the equality.

In this work a measure of identifying influential observations in testing the equality of two covariance matrices is developed using the influence curve approach. This influence measure is naturally generalized to more than two covariance matrices. A sample version of the influence measure is considered, and it includes as its component terms the well-known

[1] Departmemt of Applied Statistics, Seowon Universitiy, 360-742, KOREA.

test statistic for detecting one outlier at a time developed by Wilks (1963) and Wilks' statistic generalized to the two-group case. The influence measure can be used as a diagnostic method of detecting outliers.

2. Test of the equality of covariance matrices

Two independent random samples $\{x_1, \cdots, x_{n_1}\}$ and $\{y_1, \cdots, y_{n_2}\}$ are drawn from p -variate normal distributions $N(\mu_1, \Sigma_1)$ and $N(\mu_2, \Sigma_2)$ respectively. The Σ_i are assumed to be positive definite.

Let $\bar{x}$ and S_1 be the sample mean vector and the sample covariance matrix with divisor n_1 respectively. Similarly define $\bar{y}$ and S_2 for $N(\mu_2, \Sigma_2)$.

We consider the hypothesis of the equality of two covariance matrices:

$$H_O : \Sigma_1 = \Sigma_2 ,$$

and allow the mean vectors to vary from group to group. The alternative hypothesis is that no constraint is imposed on the whole parameters. Under both hypotheses the maximum likelihood estimators of μ_1 and μ_2 are $\bar{x}$ and $\bar{y}$ respectively, and that of the common covariance matrix under H_0 is $r_1 S_1 + r_2 S_2$, where $r_i = n_i / n_+$ $(i = 1,2)$ and $n_+ = n_1 + n_2$. Under the alternative hypothesis the maximum likelihood estimators of Σ_i are the S_i. Hence it is easily shown that the likelihood ratio statistic for testing H_0 is given by

$$\left\{ \frac{|S_1|^{r_1} |S_2|^{r_2}}{|r_1 S_1 + r_2 S_2|} \right\}^{n+/2}$$

(for more details, refer to Seber, 1984). As a test criterion, it is thus enough to take

$$|S_1|^{r_1} |S_2|^{r_2} \ / \ |r_1 S_1 + r_2 S_2|.$$

3. Influence curve for the likelihood ratio criterion

For the purpose of detecting influential observations in estimating parameters, a common way is to use the influence curves for those parameters. Let $\theta = \theta(F_1, F_2)$ be a parameter expressed as a functional of F_1 and F_2. The influence curve for θ is determined by perturbing only one of the distribution functions. The perturbation of the distribution function F_1 at z can be written as $F_1(\varepsilon) = (1-\varepsilon)F_1 + \varepsilon\delta_z$ for $0 \leq \varepsilon \leq 1$, where δ_z denotes the distribution having unit mass at the point z. The influence curve for θ at z (Hampel, 1974), when F_1 is perturbed and F_2 is kept unchanged, is defined by

$$\lim_{\varepsilon \to 0} \frac{\theta(F_1(\varepsilon), F_2) - \theta}{\varepsilon}$$

and it measures the instantaneous rate of change of θ as F_1 moves infinitesimally towards δ_z. A large absolute value would mean that the point z has a large influence in estimating the parameter $\theta(F_1, F_2)$. If $\theta(F_1(\varepsilon), F_2)$ can be expanded in a Taylor series of ε, then the influence curve for θ at z is the coefficient of the first order ε-term in a series expansion of $\theta(F_1(\varepsilon), F_2)$. The other case is similarly defined.

Before trying to get the desired influence curve, we need to define an appropriate functional which can reflect the effect of a perturbation on the likelihood ratio statistic. Let $\mu = \mu(F)$ and $\Sigma = \Sigma(F)$ be functionals of the distribution function F representing the mean vector and the covariance matrix respectively. We denote by F_i the distribution $N(\mu_i, \Sigma_i)$ for $i = 1,2$. Then $\mu_i = \mu(F_i)$ and $\Sigma_i = \Sigma(F_i)$. Let $\Sigma_H = \Sigma_H(F_1, F_2)$ be the functional of the distribution functions F_i defined by

$$\Sigma_H = \pi_1 \Sigma_1 + \pi_2 \Sigma_2,$$

where $\pi_1 + \pi_2 = 1$ and $\pi_i > 0$ for $i = 1,2$. Here each π_i represents a population proportion and is estimated by r_i for sample versions of the influence curve. The functional Σ_H reduces to the common covariance matrix whenever the null hypothesis H_0 holds. Define a functional

$$\theta(F_1,F_2) = \frac{|\Sigma_1|^{\pi_1}|\Sigma_2|^{\pi_2}}{|\Sigma_H|}.$$

When the π_i are replaced by the r_i, the value of $\theta(F_1,F_2)$ at the empirical distribution functions yields the likelihood ratio criterion as in Section 2, and therefore $\theta(F_1,F_2)$ is a reasonable functional for performing the influence analysis of the likelihood ratio statistic.

Only terms of order ε will be retained in the computation of the influence curve since the influence curve is determined by the term of order ε. To get the perturbation of θ, we derive the perturbations of its numerator and denominator separately and then combine the two perturbations. First we consider the case in which F_1 is perturbed at z and F_2 kept unchanged. The perturbation of the numerator of θ is entirely determined by that of Σ_1. The perturbation $\Sigma(F_1(\varepsilon))$ of Σ_1 becomes

$$\Sigma(F_1(\varepsilon))=\Sigma_1+\varepsilon\ \{(z-\mu_1)(z-\mu_1)^T-\Sigma_1\}+O(\varepsilon^2).$$

The determinant of $\Sigma(F_1(\varepsilon))$ can be written as

$$|\Sigma(F_1(\varepsilon))| = |\Sigma_1|\ |I_p+\varepsilon\Sigma_1^{-1}\{(z-\mu_1)(z-\mu_1)^T-\Sigma_1\}|.$$

Since $|I_p+W| = \Pi_{i=1}^{p}(1+\gamma_i)$, where the γ_i are the eigenvalues of W, $|I_p+\varepsilon\Sigma_1^{-1}\{(z-\mu_1)(z-\mu_1)^T-\Sigma_1\}|$ is equivalent to $\Pi_{i=1}^{p}(1+\varepsilon a_i)$, where the a_i are the eigenvalues of $\Sigma_1^{-1}\{(z-\mu_1)(z-\mu_1)^T-\Sigma_1\}$ and it is a polynomial in ε of order p. For a sufficiently small ε, Taylor series expansion for $|I_p+\varepsilon\Sigma_1^{-1}\{(z-\mu_1)(z-\mu_1)^T-\Sigma_1\}|$ is $1+\varepsilon\Sigma_{i=1}^{p} a_i+O(\varepsilon^2)$ and $\Sigma_{i=1}^{p} a_i$ is just the trace of $\Sigma_1^{-1}\{(z-\mu_i)(z-\mu_1)^T-\Sigma_1\}$. After applying Taylor series expansion once again, we get

$$|\Sigma(F_1(\varepsilon))|^{\pi_1} = |\Sigma_1|^{\pi_1}[1+\varepsilon\pi_1\{(z-\mu_1)^T\Sigma_1^{-1}(z-\mu_1)-p\}]+O(\varepsilon^2).$$

The perturbation of Σ_H is

$$\Sigma_H(F_1(\varepsilon),F_2)=\Sigma_H+\varepsilon\pi_1\{(z-\mu_1)(z-\mu_1)^T-\Sigma_H\}+O(\varepsilon^2).$$

For a sufficiently small ε, $(I_p+\varepsilon W)^{-1} = I_p-\varepsilon W$ and thus a Taylor series expansion yields

$$\Sigma_H(F_1(\varepsilon),F_2)^{-1} = [I_p+\varepsilon\pi_1\Sigma_H^{-1}\{\Sigma_H-(z-\mu_1)(z-\mu_1)^T\}]\ \Sigma_H^{-1}+O(\varepsilon^2)$$

from which an expansion for $|\Sigma_H(F_1(\varepsilon),F_2)|^{-1}$ becomes

$$|\Sigma_H(F_1(\varepsilon),F_2)|^{-1} = |\Sigma_H|^{-1}(1+\varepsilon\pi_1\sum_{i=1}^{p}\lambda_i) + O(\varepsilon^2),$$

where the λ_i are the eigenvalues of $\Sigma_H^{-1}\{\Sigma_H-(z-\mu_1)(z-\mu_1)^{\mathrm{T}}\}$. Note that the sum of the eigenvalues λ_i can be expressed as

$$\sum_{i=1}^{p}\lambda_i = p-(z-\mu_1)^{\mathrm{T}}\Sigma_H^{-1}(z-\mu_1).$$

An expansion for $\theta(F_1(\varepsilon),F_2)$ is thus given by

$$\theta(F_1(\varepsilon),F_2) = \theta(F_1,F_2)\{1+\varepsilon\pi_1(z-\mu_1)^{\mathrm{T}}(\Sigma_1^{-1}-\Sigma_H^{-1})(z-\mu_1)\}+O(\varepsilon^2)$$

from which the influence curve for $\theta(F_1,F_2)$ is obvioulsy determined. The influence curve includes $\pi_1\theta(F_1,F_2)$ as a multiplicative factor which is redundant for analyzing the influence of observations using its sample version. As an influence measure, it suffices to consider

$$(z-\mu_1)^{\mathrm{T}}(\Sigma_1^{-1}-\Sigma_H^{-1})(z-\mu_1)$$

which will be denoted by $IM_1(\theta(F_1,F_2), \mathbf{z})$. This influence measure is proportional to the influence curve divided by the functional of interest and it can be considered as representing a relative rate of change of $\theta(F_1,F_2)$ due to the perturbation of F_1.

By the symmetric role of the distributions in the functional $\theta(F_1,F_2)$, the influence measure, when F_2 is perturbed at $\mathbf{z}$, can be easily found as

$$IM_2(\theta(F_1,F_2), \mathbf{z}) = (z-\mu_2)^{\mathrm{T}}(\Sigma_2^{-1}-\Sigma_H^{-1})(z-\mu_2).$$

The process of deriving the influence measure in the above is easily extended to the case of more than two covariance matrices.

4. A Sample version

Three sample versions of the influence curve may be considered as in Critchley (1985): the empirical influence curve, the deleted empirical influence curve and the sample influence curve. The last two needs much more computation than the first in our case. However, the exact numerical computation is possible for the sample influence curve. Here we consider only the empirical version of $IM_i(\theta(F_1, F_2), \mathbf{z})$ which is based on the empirical distribution functions and include the definition of the sample influence curve to compare both sample versions numerically in Section 5.

Let $\widehat{F}_i$ be the empirical distribution function based on the corresponding random sample of size n_i. Then we have $\mu(\widehat{F}_1) = \overline{\mathrm{x}}$,and $\Sigma(\widehat{F}_1) = S_1$, and similar results for $\widehat{F}_2$. The empirical influence measure at x_m is obtained by substituting the $\widehat{F}_i$ for the F_i and x_m for $\mathbf{z}$, and it becomes

$$IM_1(\theta(\widehat{F}_1, \widehat{F}_2), \mathrm{x}_m) = (\mathrm{x}_m - \overline{\mathrm{x}})^T \{S_1^{-1} - (r_1 S_1 + r_2 S_2)^{-1}\} (\mathrm{x}_m - \overline{\mathrm{x}}).$$

The empirical influence measure includes interesting terms. The term $(\mathrm{x}_m - \overline{\mathrm{x}})^T S_1^{-1} (\mathrm{x}_m - \overline{\mathrm{x}})$ has been used as a test statistic for detecting outliers (Wilks, 1963). This is also equivalent to the likelihood ratio statistic for the mean slippage model (for the definition of mean slippage model, refer to Caroni and Prescott, 1992). This one-sample Wilks' statistic can be viewed as a test statistic for the two-group mean slippage model, when the alternative hypothesis holds, focusing on the detection of outliers in the x-sample. The term $(\mathrm{x}_m - \overline{\mathrm{x}})^T (r_1 S_1 + r_2 S_2)^{-1} (\mathrm{x}_m - \overline{\mathrm{x}})$ can be interpreted as a statistic for detecting outliers in the two-group case when H_0 holds and our interest centers on the detection of outliers in the x-sample. This interpretation may be ascertained by deriving the likelihood ratio statistic for the two-group mean slippage model under the constraint of the equality of covariance matrices (see Appendix for its derivation). If S_1 is equal to S_2 with probabillity one, the null hypothesis absolutely holds. In this extreme case any observation does not have an effect on the likelihood ratio statistic.

In a manner similar to the above, we can get a sample version for $IM_2(\theta(F_1, F_2), \mathbf{z})$ as

$$(\mathrm{y}_m - \overline{\mathrm{y}})^T \{S_2^{-1} - (r_1 S_1 + r_2 S_2)^{-1}\} (\mathrm{y}_m - \overline{\mathrm{y}}) .$$

To investigate the effect of the x-sample, the sample influence curve for $\theta(F_1, F_2)$ at x_m can be defined by

$$SIC_1 = (n_1 - 1)\{\theta(\widehat{F}_1, \widehat{F}_2) - \theta(\widehat{F}_{1,-m}, \widehat{F}_2)\}$$

where $\widehat{F}_{1,-m}$ is the deleted version of $\widehat{F}_1$ with the mth observation x_m deleted. The sample influence curve SIC_2 for the y-sample can be formulated analogously.

5. Example

In Table 1 we have cost data on three variables for two kinds of milk transportation: 36 measurements for gasoline trucks (Group I) and 23 measurements for diesel trucks (Group II). Three variables are $x_1=$ fuel, $x_2=$ repair, $x_3=$ capital, all measured on a per-mile basis. This data set is taken from Johnson and Wichern (1992, p.276). The data for Group I was analyzed by Bacon-Shone and Fung (1987), and Caroni and Prescott (1992). Their conclusion is that it is reasonable to view observations 9 and 21 as possible outliers.

Numerical computation needed in this section is carried out using Splus on IBM PC. The sample mean vectors are $\overline{\mathrm{x}}$ = (12.22, 8.11, 9.59) and $\overline{\mathrm{y}}$ = (10.11, 10.76, 18.17). The respective sample covariance matrices are

$$S_1 = \begin{pmatrix} 22.37 & 12.02 & 2.83 \\ 12.02 & 17.06 & 4.64 \\ 2.83 & 4.64 & 13.58 \end{pmatrix} \qquad S_2 = \begin{pmatrix} 4.17 & 0.73 & 2.26 \\ 0.73 & 24.73 & 7.35 \\ 2.26 & 7.35 & 44.63 \end{pmatrix}.$$

From these, we have $\theta\,(\widehat{F}_1, \widehat{F}_2) = 0.566$.

Table 2 includes some numerical results. The mth value in the second column with a heading IM shows the influence of the corresponding observation on the likelihood ratio statistic for testing the equality of two covariance matrices for each group and that multiplied by 0.345 for Group I or by 0.221 for Group II gives the value for the influence curve. The third column with a heading W_1 includes the values for the Wilks' likelihood ratio statistic for detecting a single outlier for each geoup. The values in the fourth colum with a heading W_2 are those for the Wilks' statistic generalized to the two-group case with the common covariance matrix. The fifth column with a heading SIC represents the sample influence curve values divided by 0.345 for Group I and by 0.221 for Group II so as to compare them with the empirical influence measures.

We will inspect the influences in Table 2 to get information about highly influential observations using the stem-and-leaf displays of the influences. The third and fourth columns in Group I show that observations 9 and 21 are possible outliers violating the normality assumption under both hypotheses, and the second column shows that observation 25 in addition to them is also highly influential in testing the equality of the covariance matrices. The result for the sample influence curve yields the similar conclusion. For Group II, observation 11 is highly influential against the normality under the null hypothesis, no one under the alternative hypothesis, and observation 16 in testing the equality.

This result shows that one method of detecting outliers for one purpose cannot guarantee the detection of outliers for another purpose, because an observation can be an outlier for one purpose but not for another purpose.

Appendix

In this appendix the problem of the two-group mean slippage model with the common covariance matrix is formulated and the likelihood ratio statistic for that model is derived. The two-group mean slippage model can be defined by specifying two hypotheses as follows. The null hypothesis H_0 is that $x_1,\cdots,x_{n_1}$ come from $N(\mu_1,\Sigma)$, and $y_1,\cdots,y_{n_2}$ from $N(\mu_2,\Sigma)$, and the alternative hypothesis H_1 is that one observation from the first group, x_1 say, comes from $N(\mu_1+\alpha,\Sigma)$, the others $x_2,\cdots,x_{n_1}$ from $N(\mu_1,\Sigma)$ and $y_1,\cdots,y_{n_2}$ from $N(\mu_2,\Sigma)$. Here the slippage parameter α is unknown. Let $\widehat{\Sigma_{H_i}}$ denote the maximum likelihood estimator of Σ under H_i for each $i=0,1$. Under the null hypothesis, the maximum likelihood estimators are

$$\hat{\mu}_1=\overline{x}\ ,\quad \hat{\mu}_2=\overline{y},\quad \hat{\Sigma}_{H_0}=r_1S_1+r_2S_2$$

and those under the alternative hypothesis are

$$\hat{\mu}_1=(n_1\ \overline{x}-x_1)/(n_1-1),\ \hat{\mu}_2=\ \overline{y}\ ,\ \hat{\alpha}=n_1(x_1-\overline{x})/(n_1-1)$$

$$\hat{\Sigma}_{H_1}=r_1S_1+r_2S_2-\{n_1/n_+(n_1-1)\}(x_1-\overline{x})(x_1-\overline{x})^T.$$

Under H_1, note that $\hat{\mu}_1$ is the mean for the empirical distribution function based on the random sample of size n_1-1 with x_1 deleted and that $x_1=\ \hat{\mu}_1+\hat{\alpha}$. Thus it is easy to show that the likelihood ratio statistic is a strictly increasing function of

$$\frac{|\hat{\Sigma}_{H_1}|}{|\hat{\Sigma}_{H_0}|}=1-\{n_1/n_+(n_1-1)\}(x_1-\overline{x})^T(r_1S_1+r_2S_2)^{-1}(x_1-\overline{x}).$$

Table 1. Milk transportation cost data

Group I				Group II			
No.	x_1	x_2	x_3	No.	x_1	x_2	x_3
1	16.44	12.43	11.23	1	8.50	12.26	9.11
2	7.19	2.70	3.92	2	7.42	5.13	17.15
3	9.92	1.35	9.75	3	10.28	3.32	11.23
4	4.24	5.78	7.78	4	10.16	14.72	5.99
5	11.20	5.05	10.67	5	12.79	4.17	29.28
6	14.25	5.78	9.88	6	9.60	12.72	11.00
7	13.50	10.98	10.60	7	6.47	8.89	19.00
8	13.32	14.27	9.45	8	11.35	9.95	14.53
9	29.11	15.09	3.28	9	9.15	2.94	13.68
10	12.68	7.61	10.23	10	9.70	5.06	20.84
11	7.51	5.80	8.13	11	9.77	17.86	35.18
12	9.90	3.63	9.13	12	11.61	11.75	17.00
13	10.25	5.07	10.17	13	9.09	13.25	20.66
14	11.11	6.15	7.61	14	8.53	10.14	17.45
15	12.17	14.26	14.39	15	8.29	6.22	16.38
16	10.24	2.59	6.09	16	15.90	12.90	19.09
17	10.18	6.05	12.14	17	11.94	5.69	14.77
18	8.88	2.70	12.23	18	9.54	16.77	22.66
19	12.34	7.73	11.68	19	10.43	17.65	10.66
20	8.51	14.02	12.01	20	10.87	21.52	28.47
21	26.16	17.44	16.89	21	7.13	13.22	19.44
22	12.95	8.24	7.18	22	11.88	12.18	21.20
23	16.93	13.37	17.59	23	12.03	9.22	23.09
24	14.70	10.78	14.58				
25	10.32	5.16	17.00				
26	8.98	4.49	4.26				
27	9.70	11.59	6.83				
28	12.72	8.63	5.59				
29	9.49	2.16	6.23				
30	8.22	7.95	6.72				
31	13.70	11.22	4.91				
32	8.21	9.85	8.17				
33	15.86	11.42	13.06				
34	9.18	9.18	9.49				
35	12.49	4.67	11.94				
36	17.32	6.86	4.44				

Table 2. Influence measures

Group I					Group II				
No.	*IM*	W_1	W_2	*SIC*	No.	*IM*	W_1	W_2	*SIC*
1	−0.26	1.21	1.47	−0.40	1	−1.60	2.57	4.17	−1.43
2	0.43	3.28	2.85	0.36	2	1.28	2.90	1.62	1.62
3	1.00	3.47	2.47	0.96	3	−1.64	2.86	4.50	−1.46
4	−1.01	3.36	4.37	−1.19	4	−3.33	4.92	8.25	−3.33
5	0.33	0.95	0.63	0.20	5	−4.12	7.03	11.16	−4.06
6	0.39	1.38	0.99	0.26	6	−1.19	1.58	2.77	−0.99
7	0.11	0.52	0.41	−0.02	7	2.51	3.47	0.95	2.89
8	1.12	3.28	2.16	1.07	8	0.08	0.80	0.73	0.33
9	−3.88	18.15	22.02	−2.68	9	−0.93	2.68	3.61	−0.69
10	0.04	0.12	0.07	−0.09	10	−0.80	1.81	2.62	−0.58
11	−0.40	1.07	1.47	−0.55	11	−5.08	7.66	12.74	−5.23
12	0.22	1.28	1.06	0.09	12	0.44	0.68	0.24	0.69
13	0.10	0.71	0.61	−0.03	13	−0.10	0.66	0.76	0.15
14	0.11	0.40	0.29	−0.02	14	0.43	0.60	0.17	0.68
15	1.57	4.32	2.75	1.59	15	0.48	1.52	1.04	0.74
16	0.68	2.36	1.68	0.58	16	6.02	8.25	2.23	7.5
17	0.29	1.05	0.76	0.16	17	−0.42	2.19	2.61	−0.17
18	0.78	3.12	2.34	0.71	18	−0.98	1.82	2.80	−0.77
19	0.20	0.42	0.21	0.07	19	−2.30	4.15	6.45	−2.15
20	1.89	6.53	4.64	2.14	20	−3.05	5.87	8.92	−2.86
21	−2.79	11.04	13.83	−2.92	21	1.03	2.58	1.56	1.34
22	0.22	0.51	0.29	0.08	22	0.38	0.88	0.50	0.63
23	1.56	5.24	3.68	1.65	23	−0.28	1.50	1.78	−0.34
24	0.68	1.93	1.26	0.57					
25	2.67	6.01	3.33	2.88					
26	0.67	2.35	1.68	0.57					
27	1.36	3.70	2.34	1.34					
28	0.67	1.41	0.74	0.55					
29	0.57	2.47	1.90	0.47					
30	0.24	1.79	1.55	0.12					
31	1.32	3.11	1.78	1.26					
32	0.45	2.49	2.04	0.35					
33	0.07	1.32	1.25	−0.07					
34	0.10	1.12	1.02	−0.03					
35	0.91	2.14	1.24	0.80					
36	0.78	4.40	3.63	0.77					

References

[1] Bacon-Shone, J. and W. K. Fung (1987). A New Graphical Method for Detecting Single and Multiple Outliers in Univariate and Multivariate Data, *Applied Statistics,* Vol. 36, 153-162.

[2] Campbell, N. A. (1978). The Influence Function as an Aid in Outlier Detection in Discriminant Analysis, *Applied Statistics,* Vol. 27, 251-258.

[3] Caroni, C. and Prescott, P. (1992). Sequential Application of Wilks's Multivariate Outlier Test, *Applied Statistics,* Vol. 41, 355-364.

[4] Cook, R. D. and Weisberg, S. (1982). *Residuals and Influence in Regression,* Chapman and Hall, London.

[5] Critchley, F. (1985). Influence in Principal Component Analysis, *Biometrika,* Vol. 72, 627-636.

[6] Hampel, F. R. (1974). The Influence Curve and its Role in Robust Estimation, *Journal of the American Statistical Association,* Vol. 20, 383-393.

[7] Johnson, R. A. and Wichern, D. W. (1992). *Applied Multivariate Statistical Analysis,* 3rd Ed., Prentice-Hall.

[8] Seber, G. A. F. (1984). *Multivariate Observations,* John Wiley & Sons.

[9] Wilks, S. S. (1963). Multivariate Statistical Outliers, *Sankhya,* Vol. 25, 407-426.

두개의 공분산 행렬의 동질성 검정에서의 영향치 분석

김명근[2]

요약

두개의 공분산 행렬의 동질성을 검정하는데 있어서, influence curve 방법을 이용하여 outlier를 찾는데 유용한 진단법을 제시한다. 이러한 진단법은 두개 이상의 공분산 행렬의 경우에 쉽게 일반화된다. 경험적 분포함수에 입각한 진단법의 sample version을 고려하며, 이것은 Wilks가 제안한 한개의 outlier를 찾는데 필요한 통계량과 두개의 모집단의 경우로 일반화된 Wilks 통계량를 포함한다.

[2] (360-742) 충북 청주시 모충동 231. 서원대학교 응용통계학과.