G.Ü. Fen Bilimleri Dergisi, 16(4):691-695, 2003 ISSN 1303-9709 *G.U. Journal of Science 16(4):691-695, 2003*

HOTELLING'S T^2 STATISTIC BASED ON MINIMUM-VOLUME-ELLIPSOID ESTIMATOR

Meral CANDAN ÇETİN* and Serpil AKTAŞ

Hacettepe University, Faculty of Science, Department of Statistics, 06532, Beytepe, Ankara, TURKEY, meral@hacettepe.edu.tr

ABSTRACT

Hotelling's T^2 is a popular statistic. In practice, T^2 is computed using the sample covariance matrix, S, as an estimate of the population covariance matrix, Σ. This estimator has many properties, but in some cases other covariance matrix estimators can be reasonable. In this paper, we consider alternative covariance T^2 matrix and location estimator named Minimum Volume Ellipsoid estimator for statistics in case of outliers. The simulation study was performed to compare the two statistics. Hotelling's T^2 based on minimum volume ellipsoid estimator has given better result than the classical one.

Key Words: Robust covariance matrix, minimum volume ellipsoid estimator, Hotelling's T^2, multivariate location and scale estimation, breakdown point

EN KÜÇÜK HACİMLİ ELİPS KESTİRİCİLERİNE DAYALI HOTELLING'S T^2 İSTATİSTİĞİ

ÖZET

Hotelling's T^2 popüler bir istatistiktir. Pratikte T^2 örneklem kovaryans matrisinden yararlanaılarak hesaplanır. S, kitle kovaryans matrisi Σ'nın tahmin edicisidir. Bu tahmin edicinin birçok özelliği vardır, fakat bazı durumlarda diğer kovaryans matris tahminleride sözkonusu olabilir. Bu çalışmada aykırı değer varlığında T^2 istatistiği için en küçük hacimli elips kestiricisi adı verilen alternatif bir kovaryans matrisi ve konum kestiricisi incelenmiştir. Her iki istatistiği karşılaştırmak için simulasyon çalışması yapılmıştır. Simulasyon çalışması sonucunda en küçük hacimli elips kestiricisine dayalı Hotelling's T^2'nin klasik Hotelling's T^2'den daha iyi sonuç verdiği görülmüştür.

Anahtar Kelimeler: Sağlam kovaryans matrisi, en küçük hacimli elips kestiricisi, Hotelling's T^2, çok değişkenli konum ve ölçek tahmini, bozulum noktası.

1. INTRODUCTION

It is known that some statistical methods are very sensitive to outliers (1), to reduce the influence of outliers, robust methods have been proposed (2). Robust estimates of location and scale parameters also play a very important role in statistics. Rousseeuw (3) introduced two esti-

mators, minimum-volume-ellipsoid (MVE) estimator and minimum-covariance-determinant (MCD) estimator having high breakdown point. MVE is used more widely than MCD. Since the presence of outlier may cause some problems, the effect of outlier on the distribution of Hotelling's T^2 should be investigated. We consider alternative covariance matrix and location estimator for T^2 statistics in case of outliers. Many authors have used different estimators of covariance matrix for T^2 (4). In the Second Section, classical Hotelling's T^2 was briefly introduced and the Third Section Minimum Volume Ellipsoid estimator was presented. In this study, distribution of T^2 for single sample was examined by the simulation study when outliers exist given in the Fourth Section.

2. HOTELLING'S T^2 DISTRIBUTION

The distribution of the T^2 statistics as the multivariate extension of Fisher's t for $\underline{\mu} = \underline{\mu}_0$ was derieved by Hotelling (5). Hotelling (5) introduced the statistics T^2. Denote the sample mean vector by and the sample covariance matrix by S then Hotelling's T^2 statistic is defined by,

$$T^2 = N(\bar{x} - \underline{\mu}_0)' S^{-1} (\bar{x} - \underline{\mu}_0) \qquad [1]$$

where N is the sample mean. This is an example of a quadratic form, so that T^2 is scalar.

When the null hypothesis H0: $\underline{\mu} = \underline{\mu}_0$ the following function of T^2 is F-distributed:

$$F^p_{N-p} = \frac{N-p}{p} \frac{T^2}{N-1} \qquad [2]$$

where p denotes the number of variables.

3. MINIMUM VOLUME ELLIPSOID ESTIMATORS

Outliers in multivariate data set can cause some erroneous statistical results. Several outlier detection methods have been proposed in the earlier literature.

In a multivariate sample, the aim of detecting outliers can be different. Several multivariate outlier identification rules base on robust estimators of location and scale. In recent years, Kosinski(6) proposed a new method for detection of outliers which is very resistant to high contamination of data. As classical covariance matrix is very sensitive to outliers, alternative covariance matrix have been proposed.

The minimum-volume-ellipsoid and the minimum-covariance-determinant are two of several multivariate location and scale estimators. These estimators have high finite-sample breakdown point. The use of estimators with high finite-sample breakdown point yields good performance according to masking effect.

Rousseeuw (7) introduced the affine equivariant estimator with maximal breakdown point, by putting

T(X)=center of minimal volume ellipsoid covering h points of X.

where h can be taken equal to [n+1]+1. This is called the minimum volume ellipsoid estimator (8). The covariance estimator of this is given by the ellipsoid. Because of the transform $x \rightarrow xA+b$ is an ellipsoid where A and b are the constants, MVE is an affine equivariant estimator. Such that any transformation on x does not affect the MVE.

The minimum-volume-ellipsoid estimator proposed by Rousseeuw (3) is a robust estimation of location and scale of multivariate data in the presence of outliers.

The MVE is the robust estimation of multivariate location and scale defined by minimizing the volume of an ellipsoid containing h points. These robust location and scale estimators can be used

to detect multivariate outliers and leverage points.

The MVE estimator searches for the smallest ellipse containing half of the data (9). When sampling from a multivariate normal distribution, then it rescaled these estimates so that they estimate the usual population mean and covariance matrix. It is difficult to find the smallest ellipse containing half of data. From the n points, MVE estimator randomly selects h points without replacement and computes the volume of this ellipse. The set of points giving the smallest volume is taken to be minimum volume ellipsoid. The location and scale MVE estimators yield an effective method for identifying outliers in multivariate data (10).

This estimator is defined to be the ellipsoid of minimum volume covering at least h points of the data set (8).

The breakdown point of MVE estimator at any p-dimensional sample X equals

$$\varepsilon_n^*(T,X) = (n/2 - p + 1)/n \qquad [3]$$

which converges to 50% as $n \to \infty$ (8).

Equation (1) can be redefined in terms of MVE estimators as follows,

$$T^2_{MVE} = N(\bar{x}_{MVE} - \underline{\mu}_0)' S_{MVE}^{-1} (\bar{x}_{MVE} - \underline{\mu}_0) \qquad [4]$$

In equation [4], $\bar{x}_{MVE}$ and S^{-1}_{MVE} represent the MVE estimates of the sample mean and covariance matrix.

4. SIMULATION STUDY

The simulation was performed to discuss the Hotelling's T^2 and the T^2 based on MVE estimator (T^2_{MVE}) when the data consist of outliers.

In this study, random samples having 10, 20, 50 and 150 observations and three variables were generated from multivariate normal distrubution, $N(\mu, \Sigma)$ with known mean (10, 20 and 30 respectively) and covariance matrix. In order to see the influence on the Hotelling's T^2, an outlying observations in data set were created arbitrarily in each sample. 1000 replications were performed with outlier and no outliers. Four cases were examined, as follows

- Hotelling's T^2 with no outliers (T^2)
- Hotelling's T^2 with outliers (T^2)
- T^2 based on MVE with no outliers (T^2_{MVE})
- T^2 based on MVE with outliers (T^2_{MVE})

For a given mean vector, below hypothesis was tested

$$H_0 : \underline{\mu} = \underline{\mu}_0$$

versus the alternative hypothesis,

$$H_A : \underline{\mu} \neq \underline{\mu}_0$$

Where $\underline{\mu}_0$ is the population mean vector as 10, 20 and 30 respectively. The proportions of the rejected and accepted hypothesis for 1000 replications, are given in Table 1.

In order to obtain the considered above, a program was coded using S-Plus functions.

Before the outlying observations were created, we set all the parameters the null hypothesis has been rejected.

The results show that T^2 and T^2_{MVE} are almost obtained similar in case of no outlier. It is clear that T^2 is affected by the outlying observations. T^2_{MVE} gives more successful results in the presence of outliers. Unlike T^2 with outliers, T^2_{MVE} with outliers rejected 99.3% of hypothesis for n=50 when the alternative hypothesis is true. when data consist outliers give the same proportions with T^2 when data consist of no outliers. It is determined that when the data consist of outlier, T^2_{MVE} is the most proper method.

Table 1. Simulation Results

	Ho rejected (%)		*Ho accepted (%)*	
N=10	**outlier**	**no outlier**	**outlier**	**No outlier**
T^2	14.1	96.9	85.9	3.1
T^2_{MVE}	97.2	97.4	2.8	2.6
	Ho rejected (%)		*Ho accepted (%)*	
N=20	**outlier**	**no outlier**	**outlier**	**No outlier**
T^2	6.6	100.0	93.4	0
T^2_{MVE}	100.0	100.0	0	0
	Ho rejected (%)		*Ho accepted (%)*	
N=50	**outlier**	**no outlier**	**outlier**	**No outlier**
T^2	4.3	99.4	95.7	0.6
T^2_{MVE}	99.3	99.5	0.7	0.5
	Ho rejected (%)		*Ho accepted (%)*	
N=150	**outlier**	**no outlier**	**outlier**	**no outlier**
T^2	38.6	100.0	61.4	0
T^2_{MVE}	100.0	100.0	0	0

5. CONCLUSION

As mentioned earlier, classical covariance matrix is sensitive to outliers. In this study MVE estimators as an alternative has been used. After the simulation study it is shown that the Hotelling's T^2 based on MVE estimator has given the better results than the classical estimator when data include outliers. From the results of Table 1, We could see that T^2_{MVE} gives better results for small samples (n=10, 20 and 50). For large samples and T^2 give the similar proportions when rejecting and accepting the hypothesis. Simulation study was performed for the samples over the 150, but the same results were obtained with the n=150. It is shown that when data consist of outlier, does not effect the outlier for small samples. When data consist of no outlier and T^2 are the analogous statistics. Simulation study could extent for the different parameters.

REFERENCES

1. Beckman, R.C. and Cook R.D.,"Outlier... s", ***Technometrics***, 25: 119-149 (1983).

2. Hoaglin, D. C., Mosteller F., Tukey, J.W., ***Understanding Robust and Exploratory Data Analysis***, John Wiley& Sons, Inc.,New York (1983).

3. Rousseeuw, P.J, "Multivariate Estimation with high breakdown point", ***Mathematical Statistics and Applications***, Vol. B. Reidel, Dordrecht, 283-297 (1985).

4. Chou, Y.M.,Mason, R.L., Young J.C.,"Power comparisons for a Hotelling's T^2 statistic" , ***Commun. Stat.-Simulation***, 28(4): 1031-1050 (1999).

5. Hotelling, H., "The Generalization of student's ratio", ***Annals of Mathematical Statistics***, 2: 360 (1931).
6. Konsinski, A., S., "A procedure for the detection of multivariate outliers", ***Computational Statistics & Data Analysis***, 29: 145-161 (1999).
7. Rousseeuw, P.J., "Least median of squares regression", ***Journal of the Americal Statistical Association***, 79: 871-880 (1984).
8. Rousseeuw, P.J., Leroy A.M., "Robust regression and outlier detection", ***John-Wiley***, New-York (1987).
9. Wilcox R.R, " Introduction to robust estimation and hypothesis testing", ***Academic Press*** (1997).
10. Rousseeuw, P.J, Van Zomeren, B.C., "Unmasking multivariate outliers and leverage points", ***Journal of the Americal Statistical Association***, 85: 633-639 (1990).

Received:17.06.2002 *Accepted:30.05.2003*