Multivariate Analysis and Its Applications
IMS Lecture Notes – Monograph Series (1994) Volume 24

UNBIASEDNESS OF A MULTIVARIATE OUTLIER TEST FOR ELLIPTICALLY CONTOURED DISTRIBUTIONS

BY JIAN-XIN PAN* AND XUE-REN WANG

Hong Kong Baptist College and Yunnan University

Under a class of elliptically contoured distributions, the likelihood ratio criterion (LRC) for detecting multiple outliers is established in this paper. It is shown that the LRC has the same form as in the normal case. Furthermore, the unbiasedness of the LRC is derived.

1. Introduction. Suppose that $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$ is a random sample from some p-dimensional distribution $F_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma} > \mathbf{0}$ are the mean vector and the covariance matrix, respectively, and both are unknown. We are often required to detect whether there are outliers in the sample and which sample points are outliers. This can be reduced to a testing hypothesis problem as follows. If we choose the null model in which there are no outliers in the sample

$$H : \boldsymbol{x}_i \sim F_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad i = 1, 2, \cdots n, \tag{1.1}$$

then a possible alternative model which may account for multiple outliers is the multivariate model with mean slippage

$$\begin{aligned} K : \quad & \boldsymbol{x}_i \sim F_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \qquad (i \notin I) \\ & \boldsymbol{x}_i \sim F_p(\boldsymbol{\mu} + \boldsymbol{a}_i, \boldsymbol{\Sigma}) \qquad (i \in I) \end{aligned} \tag{1.2}$$

where $I = \{i_1, i_2, \cdots, i_k\}$ is an index subset of $\{1, 2, \cdots, n\}$ with a fixed positive integer k and $\boldsymbol{a}_{i_1}, \cdots, \boldsymbol{a}_{i_k}$ are unknown mean slippage parameters.

For hypotheses (1.1) and (1.2), Siotaui (1959) and Wilks (1963) discussed the likelihood ratio criterion under the multivariate normal distribution $F_p = N_p$. In this paper, we extend their results to a class of elliptically contoured distributions. The null distribution of the likelihood ratio testing statistic for detecting the multiple outliers is shown to be a Wilk's distribution, which has the same distribution as in the normal case. Furthermore, the unbiasedness of the likelihood ratio test is derived. Under the assumption of an elliptically contoured distribution, Sinha (1984) provided a locally best invariant test for outliers based on multivariate sample kurtosis, which is useful for identifying

---

* Supported partially by the WAI TAK Investment and Loan Company LT.D. Research Scholarship of Hong Kong.

AMS 1991 Subject Classifications: Primary 62H15; Secondary 62H10.

Key words and phrases: Elliptically contoured distribution; likelihood ratio criterion; multiple outlier; unbiasedness of a test.

whether there are outliers in the sample. The likelihood ratio test for discordantly multiple outliers, however, can be used not only for determining whether there are outliers in the sample, but also for detecting which of the observations are discordant outliers.

2. Likelihood Ratio Criteria. Suppose that the characteristic function of a random matrix $\boldsymbol{X}_{n\times p} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n)^\tau$ has the form:

$$exp\Big\{i\sum_{j=1}^{n} \boldsymbol{t}_j^\tau \boldsymbol{\mu}_j\Big\} \cdot \Phi\Big(\sum_{j=1}^{n} \boldsymbol{t}_j^\tau \boldsymbol{\Sigma}_j \boldsymbol{t}_j\Big), \tag{2.1}$$

then $\boldsymbol{X}$ is known to have a matrix elliptically contoured distribution (See, e.g., Fang and Zhang, 1990). When there is a density function of $\boldsymbol{X}$, it must be of the form:

$$\prod_{j=1}^{n} |\boldsymbol{\Sigma}_j|^{-1/2} g\Big(\sum_{j=1}^{n} tr\boldsymbol{\Sigma}_j^{-1}(\boldsymbol{x}_j - \boldsymbol{\mu}_j)(\boldsymbol{x}_j - \boldsymbol{\mu}_j)^\tau\Big), \tag{2.2}$$

where $g(\cdot)$ is a non-negative function and satisfies $\int_0^{+\infty} y^{\frac{np}{2}-1} g(y)dy < +\infty$. In this case we denote $\boldsymbol{X} \sim LEC_{n\times p}(\boldsymbol{M};\ \boldsymbol{\Sigma}_1,\ \boldsymbol{\Sigma}_2, \cdots, \boldsymbol{\Sigma}_n;\ g)$, where $\boldsymbol{M} = (\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots, \boldsymbol{\mu}_n)^\tau$. Especially, if $\boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_n = \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}_1 = \boldsymbol{\Sigma}_2 = \cdots = \boldsymbol{\Sigma}_n = \boldsymbol{\Sigma}$ (say) we denote $\boldsymbol{X} \sim LEC_{n\times p}(\boldsymbol{1_n}\boldsymbol{\mu}^\tau;\ \boldsymbol{I_n} \otimes \boldsymbol{\Sigma};\ g)$, where $\boldsymbol{1_n} = (1, \cdots, 1)^\tau \in R^n$ and the notation $\otimes$ represents the Kronecker product of two matrices. In this case, $\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n$, the rows of the matrix $\boldsymbol{X}$, are mutually uncorrelated and have a common mean vector $\boldsymbol{\mu}$ and a covariance matrix $\boldsymbol{\Sigma}$. Now, the tests (1.1)–(1.2) in Section 1 become

$$H : \boldsymbol{X} \sim LEC_{n\times p}(\boldsymbol{1_n}\boldsymbol{\mu}^\tau;\ \boldsymbol{I_n} \otimes \boldsymbol{\Sigma};\ g) \tag{2.3}$$

and

$$K : \boldsymbol{X} \sim LEC_{n\times p}(\boldsymbol{1_n}\boldsymbol{\mu}^\tau + \boldsymbol{DA};\ \boldsymbol{I_n} \otimes \boldsymbol{\Sigma};\ g) \tag{2.4}$$

where $\boldsymbol{D} = (\boldsymbol{e}_{i_1}, \boldsymbol{e}_{i_2}, \cdots, \boldsymbol{e}_{i_k})$ and $\boldsymbol{A} = (\boldsymbol{a}_{i_1}, \boldsymbol{a}_{i_2}, \cdots, \boldsymbol{a}_{i_k})^\tau$ are $n \times k$ and $k \times p$ matrices, respectively, $\boldsymbol{e}_i$ $(i \in I)$ is a n-dimensional vector whose $i-$th element is one and the remaining elements are zeros, and $\boldsymbol{a}_i$ $(i \in I)$ is an unknown mean slippage p-variate vector parameter.

THEOREM 2.1. *Suppose the function $g(y)$ in (2.2) is continuous and decreasing for $y \in [0, +\infty)$, and $n > p + k + 1$. Then the likelihood ratio criterion of the tests (2.3)–(2.4) is equivalent to rejecting H if*

$$\Lambda_I = |\boldsymbol{S}_{(I)}|/|\boldsymbol{S}| \le C_\alpha \tag{2.5}$$

where C_α is the lower $100\alpha\%$ critical point of the Wilks' distribution $\Lambda(p, n-k-1, k)$, $\overline{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{x}_i$, $\boldsymbol{S} = \sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\tau$, $\overline{\boldsymbol{x}}_{(I)} = \frac{1}{n-k}\sum_{i\notin I} \boldsymbol{x}_i$ and $\boldsymbol{S}_{(I)} = \sum_{i\notin I}(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(I)})(\boldsymbol{x}_i - \overline{\boldsymbol{x}}_{(I)})^\tau$.

PROOF. It can be shown that the likelihood ratio testing statistic is equivalent to $\Lambda_I = |\boldsymbol{S}_{(I)}|/|\boldsymbol{S}|$. In the following only the null distribution of Λ_I is derived. Let $\Lambda_I(\boldsymbol{X})$ denote the statistic Λ_I for the data matrix $\boldsymbol{X}$. Under the null hypothesis H, it is obvious that $\Lambda_I(\boldsymbol{X} - \mathbf{1}_n\boldsymbol{\mu}^\tau) \equiv \Lambda_I(\boldsymbol{X})$ and $\Lambda_I(\alpha\boldsymbol{X}) \equiv \Lambda_I(\boldsymbol{X})$ are true for all $\boldsymbol{\mu} \in R^p$ and $\alpha > 0$, respectively. Therefore, the null density distribution of Λ_I does not depend on the choice of the function $g(y)$ (See, e.g., Fang and Zhang, 1990). In other words, it is distribution-free or distribution-robust on the class of elliptically contoured distributions. Especially, the null density distribution of Λ_I is the same that of Λ_I as $g(y) = (2\pi)^{-pn/2}e^{-y/2}$ $(y \geq 0)$, which corresponds to the matrix normal distribution $\boldsymbol{X} \sim N_{n,p}(\mathbf{1}_n\boldsymbol{\mu}^\tau, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$. Noting that $\boldsymbol{S} = \boldsymbol{X}^\tau\boldsymbol{D}_n\boldsymbol{X}$ and $\boldsymbol{S}_{(I)} = \boldsymbol{X}^\tau\boldsymbol{C}_1\boldsymbol{X}$ where $\boldsymbol{D}_n = \boldsymbol{I}_n - \frac{1}{n}\mathbf{1}_n\mathbf{1}_n^\tau$, $\boldsymbol{C}_1 = diag(\boldsymbol{D}_{n-k}, \mathbf{0})$ and $\boldsymbol{X} = (\boldsymbol{X}_{(I)}^\tau, \boldsymbol{X}_I^\tau)^\tau$, we know that $\boldsymbol{C}_1^2 = \boldsymbol{C}_1$, $\boldsymbol{C}_1^\tau = \boldsymbol{C}_1$, $rk(\boldsymbol{C}_1) = n - k - 1$ and $\boldsymbol{C}_1(\mathbf{1}_n\boldsymbol{\mu}^\tau) = \mathbf{0}$ so that $\boldsymbol{S}_{(I)} \sim W_p(n - k - 1, \boldsymbol{\Sigma})$ under the multivariate normal assumption. Let $\boldsymbol{C}_2 = \boldsymbol{D}_n - \boldsymbol{C}_1$ and $\boldsymbol{E}_{(I)} = \boldsymbol{S} - \boldsymbol{S}_{(I)}$ then $\boldsymbol{E}_{(I)} = \boldsymbol{X}^\tau\boldsymbol{C}_2\boldsymbol{X}$. It can be verified that $\boldsymbol{C}_2^2 = \boldsymbol{C}_2$, $\boldsymbol{C}_2^T = \boldsymbol{C}_2$, $\boldsymbol{C}_1\boldsymbol{C}_2 = \mathbf{0}$, $rk(\boldsymbol{C}_2) = k$ and $\boldsymbol{C}_2(\mathbf{1}_n\boldsymbol{\mu}^\tau) = \mathbf{0}$. Therefore when $\boldsymbol{X} \sim N_{n,p}(\mathbf{1}_n\boldsymbol{\mu}^\tau, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma})$, $\boldsymbol{E}_{(I)} = \boldsymbol{S} - \boldsymbol{S}_{(I)}$ is independent of $\boldsymbol{S}_{(I)}$ and $\boldsymbol{E}_{(I)} \sim W_p(k, \boldsymbol{\Sigma})$. Furthermore,

$$\Lambda_I = |\boldsymbol{S}_{(I)}|/|\boldsymbol{S}_{(I)} + \boldsymbol{E}_{(I)}| \sim \Lambda(p, n - k - 1, k) \tag{2.6}$$

under the null hypothesis H. The proof is complete.

REMARK 2.1. *When $I = \{i\}$ $(i = 1, 2, \cdots, n)$, the $i-th$ observation is declared as a discordantly single outlier of size α if*

$$T_i^2 \geq \frac{p(n-1)C_\alpha^*}{n(n-p-1) + npC_\alpha^*} \tag{2.7}$$

where $T_i^2 = (\boldsymbol{x}_i - \overline{\boldsymbol{x}})^\tau\boldsymbol{S}^{-1}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})$ and C_α^* is the upper $100\alpha\%$ critical point of $F_{(p,n-p-1)}$. When $I = \{i, j\}$ $(i, j = 1, 2, \cdots, n)$, the $(i, j)-th$ observation is declared as a discordant outlier pair of size α if

$$\sqrt{\Lambda_{ij}} \leq \frac{n-p-3}{pC_\alpha^{**} + (n-p-3)} \tag{2.8}$$

where C_α^{**} is the upper $100\alpha\%$ critical point of $F_{(2p,2(n-p-3))}$.

REMARK 2.2. *In most practical problems, the index subset I can not be given in advance. Suppose k is fixed, to detect a set of k outliers, a reasonable testing statistic is $\Lambda_{\min}^k = \min_I\{\Lambda_I\}$ where I runs over all subsets including k indexes. Unfortunately, the exact null distribution of $\Lambda_{\min}^k$ is unknown. In this case, Bonferroni's principle in multiple comparisons can be recommended.*

THEOREM 2.2. *Suppose the function $g(y)$ in (2.2) is continuous and decreasing for $y \in [0, +\infty)$, and $n > p + k + 1$. Then the likelihood ratio test (2.5) for detecting multiple outliers of a sample in elliptically contoured distribution is unbiased.*

PROOF. Obviously, the power function of the test (2.5) $\Pr\{\Lambda_I \leq C_\alpha\}$ is a function of the mean shift parameter $\boldsymbol{A}$. It can be shown, however, $\Pr\{\Lambda_I \leq C_\alpha\}$ depends upon $\boldsymbol{A}$ only through $\|\boldsymbol{\delta}_i^*\| = \sqrt{\boldsymbol{\delta}_i^{*\tau}\boldsymbol{\delta}_i^*}$ $(i = 1, 2, \cdots, k)$ so that we denote it as $\beta(\|\boldsymbol{\delta}_1^*\|, \|\boldsymbol{\delta}_2^*\|, \cdots, \|\boldsymbol{\delta}_k^*\|)$, where $\boldsymbol{\delta}_i^{*\tau}$ is the $i-$th row of $\boldsymbol{\Delta}^* = \boldsymbol{\Gamma}\boldsymbol{A}\boldsymbol{\Sigma}^{-1/2}$ and $\boldsymbol{\Gamma}$ is a $k \times k$ orthogonal matrix with the first row $\mathbf{1}_k^\tau/\sqrt{k}$. Furthermore, $\beta(\|\boldsymbol{\delta}_1^*\|, \|\boldsymbol{\delta}_2^*\|, \cdots, \|\boldsymbol{\delta}_k^*\|)$ can be shown to be a monotonically increasing function of $\|\boldsymbol{\delta}_i^*\|$ $(i = 1, 2, \cdots, k)$, thus $\beta(\|\boldsymbol{\delta}_1^*\|, \|\boldsymbol{\delta}_2^*\|, \cdots, \|\boldsymbol{\delta}_k^*\|) \geq \beta(0, 0, \cdots, 0)$, which implies that the power function of the test (2.5) achieves its minimum at the null hypothesis $H : \boldsymbol{A} = \mathbf{0}$. In other words, the likelihood ratio test (2.5) is unbiased and the proof is complete.

Acknowledgement. The authors would like to thank Professor K.T. Fang and Dr. F.J. Hickernell for their valuable suggestions and discussions. Also, the referees' helpful comments are appreciated.

REFERENCES

FANG, K. T. and ZHANG, Y. T. (1990). *Generalized Multivariate Analysis.* Springer–Verlag and Science Press, Berlin and Beijing.

SINHA, B. K. (1984). Detection of multivariate outliers in elliptically symmetric distributions. *Ann. of Stat.* **12**, 4, 1558–1565.

SIOTANI, M. (1959). The extreme value of the generalized distances of the individual points in the multivariate normal sample. *Ann. Inst. Stat. Math.*, Tokyo **10**, 10, 183–208.

WILKS, S. S. (1963). Multivariate statistical outliers. *Sankhya A* **25**, 406–427.

DEPARTMENT OF MATHEMATICS
HONG KONG BAPTIST COLLEGE
224 WATERLOO ROAD, KOWLOON
HONG KONG

DEPARTMENT OF STATISTICS
YUNNAN UNIVERSITY
KUNMING 650091, CHINA