Mathematical Theory and Modeling
ISSN 2224-5804 (Paper) ISSN 2225-0522 (Online)
Vol.8, No.8, 2018

A Modified Approach to Identifying Multiple Outliers in Multivariate Data

Benony Kwaku Gordor

Department of Computer Science, Ashesi University, 1 University, Avenue, Berekuso, PMB CT 3 Cantonments, Accra, Ghana

E-mail of the corresponding author: bgordor@ashesi.edu.gh

Abstract

Using the likelihood-based Wilks' method to identify multiple outliers in multivariate datasets can be computationally intensive. For example, to identify the most extreme *k* outliers in a sample of size *n*, the likelihood principle requires that we examine all $\binom{n}{k}$ subsets of observations. Faced with this amount of computation, it seems reasonable to suggest that one examines only observations that lie at the periphery of the data 'cloud' rather than all the observations. However, this 'short-cut' approach raises two major questions. The first is how many observations at the periphery one has to examine and second, what the guarantee is that the observations that are identified are truly the outliers. This paper is an attempt to provide answers to these questions by conducting a simulation study involving a large number of datasets of various sizes, dimensions and degree of outlier contamination. In the end, a formula for identifying multiple outliers based on the approach is developed.

Key words: Outliers, Mahalanobis Distance, Optimal Subset, Failure Rate, Risk Table

1. Introduction

Practical problems in the identification of multiple outliers in multivariate datasets usually make use of the Wilks' ratio statistic given by

$$R_T = \frac{\left|\mathbf{S}_{(T_k)}\right|}{\left|\mathbf{S}\right|}, \qquad (1)$$

which determines the subset T_k of *k* observations among the sample of size *n* for which the ratio is minimum, where $\mathbf{S}$ is the sum of squares and cross-product matrix, and $\mathbf{S}_{(T_k)}$ is the corresponding matrix with the observations in T_k removed from the sample. The problem also uses any other method based on the likelihood ratio principle. This principle requires a two-stage approach to outlier identification. The principle as constructed by Barnett (1979) states that given a basic model, *F*, and an alternative contaminating model, *F'*, the most extreme observation is that one, x_i, whose assignment as the contaminant in the sense of *F'* maximizes the difference between the log-likelihoods of the sample under F' and F. If this difference is surprisingly large, declare x_i to be an outlier. The main concern in this paper is the application of the first stage of this principle because of the computational difficulties involved.

In order to find the most extreme *k* observations, the first stage of the principle requires that we determine which

subset of size *k* maximizes the log-likelihood under F'. This means that in a sample of size *n*, we must evaluate the log-likelihood for $\binom{n}{k}$ subsets. If we use the Wilks' ratio method, for example, it means that we must calculate $\binom{n}{k}$ *k*-outlier scatter-ratios and examine them for extremeness. In large multi-dimensional datasets, this can be computationally prohibitive, even for a moderate value of *k*. For example, for even $n = 40$ and $k = 2$, there are 780 ratios to examine; and for $n = 50$ and $k = 3$, there are 19,600 such ratios to examine. It is obviously prohibitive to examine all of these ratios.

In univariate data, we could order the *n* observations as $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$, so that the most outlying *k*-tuple is clearly a subset of the first few elements on either side. We may, for example, choose this subset to contain the lowest *k* and the highest *k* values:

$$x_{(1)} < x_{(2)} < \cdots < x_{(k)}, x_{(n)}, x_{(n-1)}, \ldots, x_{(n-k)},$$

so we need to examine at most $\binom{2k}{k}$ subsets. Figure 1 shows a simple one-dimensional scatter-plot of a dataset credited to Chauvenet (Barnett and Lewis, 1994). The data relates to 15 observations of the vertical semi-diameter of Venus made by Herndon in 1846. If we have to test for $k = 2$ outliers, for example, based on the likelihood principle, we have to examine all the $\binom{15}{2}$ for extremeness.

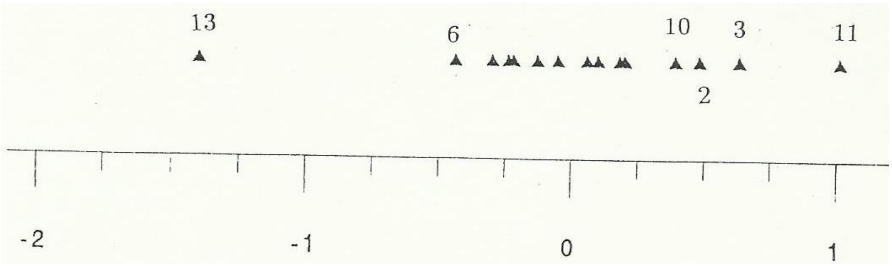

Figure 1: One dimensional display of semi-diameters of Venus

However, it is easy to see from the diagram that six observations, 13, 6, 10, 2, 3, and 11 are reasonably separated from the rest of the observations and can be considered candidate outliers. Therefore, we may consider examining only $\binom{6}{2}$ rather than all the $\binom{15}{2}$ subsets.

By extension to the multivariate case, the *p*-dimensional $(p > 1)$ dataset is reduced to a univariate dataset. A dimensional reduction technique employed in this case is the Mahalanobis squared generalized distance of each observation from their mean given by

$$U_n = (\mathbf{x}_i - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_i - \bar{\mathbf{x}}). \qquad (2)$$

Subsequently, the reduced observations are ordered in order of magnitude of their distances, then examine $\binom{m}{k}$, where $m (< n)$ is the number of observations with largest distances for outliers.

In a simulated study, Gordor (1994) generated a large number of datasets of different sizes and dimensions. For each dataset, he calculated a failure rate, which he defined as the number of times that m_k observations with largest Mahalanobis squared distance failed to contain the most extreme *k* outliers. The study showed that

providing that the sample size is reasonably large $(n > 30)$, failure rate is only weakly affected by the dimension of the data. The objective of this paper is to determine the number of observations, m_k, such that examining only $\binom{m_k}{k}$ results in the *k* outliers being identified with 0% failure rate. We call this the optimum subset of extreme points.

2. Method

In this paper we shall determine the optimum subset for $k = 1, 2, 3$ and 4 extreme points or outliers. This work is limited to a maximum of 4, with the understanding that extreme values beyond 4 may not strictly be a problem of outlier identification but a discrimination between (two) groups. Given that failure rate is only weakly affected by the dimension of the data for large $(n > 30)$ sample size, we generated a sample of size $n = 50$ and dimensionality of $p = 4$ in each case.

2.1 Unknown Number of Outliers

For this purpose, 100 datasets are generated each according to the following rules: With probability 0.9, sample from $N_p(0, 1)$ and with probability 0.1, sample from $N_p(c, 1)$. Computationally, this means we generate *x* from $N_p(0, 1)$ and generate *u* from the uniform distribution $U(0, 1)$. If $u > 0.9$, we add *c* to *x*; if $u < 0.9$, we do not add anything. The constant *c* can be any number. In this study, we chose $c = 1$ for the first 20 datasets, $c = 3$ for the next 30 sets, and $c = 5$ for the remaining 50 sets. This strategy is to ensure that we have three potential sources of outlier contamination. Table 1 shows failure rates when *m* observations with the largest distances are examined for the most extreme *k* outliers.

Table 1: Failure rate (%) in the case of unknown outliers

No. of observations examined, *m*	Single	Pair	Triple	Q'duple
1	0	-	-	-
2	0	39	-	-
3	0	15	84	-
4	0	7	65	100
5	0	3	49	98
6	0	2	39	97
7	0	0	35	94
8	0	0	31	90

2.2 Two Known Outliers in Each Dataset

Dataset of size 50 are generated from $N_4(0, 1)$ as before. Then we introduced two discordant outliers by adding two suitably chosen constants randomly to two of each of the 50 observations. (Two outliers are judged discordant at 5% significant level when their Wilks' ratio r_2 is such that $\sqrt{r_2} < 0.688$. The process was repeated until 100 datasets, each contaminated with two discordant outliers, were obtained. Table 2 shows failure rates when *m* observations with the largest distances are examined for the most extreme *k* outliers.

Table 2: Failure rate (%) for a pair of outliers

No. of observations examined, *m*	2	3	4
Failure rate (%)	3	1	0

It can be observed from the table that in 3% of the samples, the process fails to identify the two outliers. In other words, failure rate for samples contaminated with two outliers is 3% when we limit the search to the two most extreme values; and if we extend the search for these outliers to three most extreme values, the failure rate falls to 1%. Finally, when the search for the outliers is extended to four most outlying observations, the failure rate is absolutely 0%. We can therefore conclude that to identify two outliers in a dataset, it is safe to search for them among the four most extreme observations only. Computationally, it means we need to search for a pair of outliers by examining only $\binom{4}{2} = 6$ most extreme observations.

2.3 Three Known Outliers in each Dataset

We again generated data sets of size 50 each from $N_4(0, 1)$ as before. Then we introduced three discordant outliers by adding three suitably chosen constants randomly to three of each of the 50 observations. The process was repeated until 100 datasets, each contaminated with three discordant outliers, were obtained. Table 3 shows failure rates when *m* observations with the largest distances are examined for the most extreme *k* outliers.

Table 3: Failure rate (%) for triple outliers

No. of observations examined, *m*	3	4	5	6	7
Failure rate (%)	46	20	3	1	0

It can be observed from the table that in 46% of the samples, the process fails to identify the three outliers. In other words, failure rate for samples contaminated with three outliers is 46% when we limit the search to the three most extreme values; and if we extend the search for these outliers to four most extreme values, the failure rate falls to 20%. Finally, when the search for the outliers is extended to seven most outlying observations, the failure rate is absolutely 0%. We can therefore conclude that to identify thee outliers in a data set, it is safe to search for them among the seven most extreme observations only. Computationally, it means we need to search for a triple of outliers by examining only $\binom{7}{3} = 35$ most extreme observations.

2.4 Three Known Outliers in each Dataset

A similar simulation study shows (Table 4) that in the case of four outliers, we need to search among the nine most extreme observations.

Table 4: Failure rate (%) for quadruple outliers

No. of observations examined, *m*	4	5	6	7	8	9
Failure rate (%)	91	63	30	16	8	0

Searching for four outliers among the four or five most extreme points is prone to large failure rate.

3. Results and Discussion

3.1 Optimum Number of Extreme Observations to Examine for Outliers

The results of the simulation exercise displaced in Tables 1 to 4 are summarized in Table 5. We refer to this table as 'Risk Table' for identifying multiple outliers in a multivariate data set. In the table, the figures represent failre rates when m observations with the largest Mahalanobis squared distances are examined for *k*-tuple (k = 1, 2, 3, 4) outliers.

Table 5: Risk table for finding multiple outliers in multivariate data

	No. of observations examined, m_k								
No. of outliers, *k*	1	2	3	4	5	6	7	8	9
1	0	0	0	0	0	0	0	0	0
2	-	3	1	0	0	0	0	0	0
3	-	-	46	20	3	1	0	0	0
4	-	-	-	91	63	30	16	8	0

From the table, it can be observed that the optimum number, m_k, of most extreme observations that guarantees the correct choice of two discordant outliers is 4; that for three outliers is 7; and that for four outliers is 9. A striking relationship between m_k, and k is now clear. This is,

$$m_k = \begin{cases} 2k, & k = 2 \\ 2k+1, & k > 2 \end{cases}$$

The relationship above therefore provides the following simple rule for identifying multiple outliers in multivariate datasets. The rule states that:

> To find the *k*-tuple of discordant outliers in a multivariate dataset, one needs to examine at most $2k+1$ most extreme observations, as ordered by their generalized squared distances.

3.2 Comparing Results with Wilks' Ratio Method

We now find it appropriate to compare our modified method with the well-known Wilks' method of identifying outliers in multivariate data sets. Note that in the case of Wilks, the outliers are identified by examining all the $\binom{n}{k}$ outlier ratios rather than $\binom{m_k}{k}$, where $m_k < n$.

The data used are the well-known Iris Setosa data (Anderson, 2003) and the Transportation Cost data (Johnson & Wichern, 2002). We used these datasets because they are well-used in the literature (Wilks, 1963; Hadi, 1992; Caroni & Prescott, 1992; Nkansah & Gordor, 2012a; 2012b; 2013) on the study of outliers. The outliers identified in each case are shown in Table 6.

Table 6: Identifying Outliers in Iris Data using Wilks Ratio and Modified Method

Dataset	No. of outliers, k	Observations selected by the Wilks' Ratio Method			Modified Method		
Iris Setosa	1	42			44		
	2	42	23		44	42	
	3	42	44	23	44	42	23
Transportation Cost	1	9			9		
	2	9	21		19	21	
	3	9	21	36	9	21	36

It can be observed from the table that the two methods tend to identify the same set of outliers, except in the Iris Setosa case, where the modified method identified the single outlier as 44 as opposed to 42 by Wilks ratio method. In the case of the pair of observations, 42 is identified by both methods. However, the other member of the pair is different. For the transportation cost data, there is perfect agreement between the two methods. It is interesting to note that the observations selected by Wilks ratio method are among the $2k+1$ observations with the largest distances. Hence, it is not necessary to search for outliers by considering all the scatter ratios as in Wilks method.

4. Summary and Conclusion

The paper looked at the computational difficulties associated with identifying multiple outliers in multivariate data sets and considered a short-cut approach to the problem by simulating a large number of multivariate data sets with varying contamination levels. The approach involves examining only a subset of extreme observations determined by ordering their Mahalanobis distances for the outliers. It was found that there exists an optimum number, m_k, for every k such that when one searches among them for k-tuple outliers, a correct identification of the outliers is guaranteed. In terms of k, this optimal value is $m_k = 2k$ for two outliers, and $m_k = 2k+1$ for three or more outliers.

The performance of this modified method is then compared with Wilks' method. It was observed that the observations selected by Wilks ratio method are among the $2k+1$ observations with the largest distances. Consequently, we conclude that the amount of computation involved in using Wilks method is greatly reduced by the application of the modified method. Since it has been shown by Gordor (1994) that for large sample sizes $(n > 30)$ results are not affected by dimension and size, the modified approach can be applied to any multivariate normal datasets.

Reference

Anderson, E. (2003). *Introduction to Multivariate Statistical Analysis*. New Jersey: Prentice Hall.

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data*. John Wiley and Sons, New York.

Caroni, C., & Prescott, P. (1992). Sequential application of Wilks' multivariate outlier test. *Applied Statistics*, **41**, 355-364.

Gordor, B. K. (1994). *Some informal methods for the detection and display of outliers in data*, Ph.D. Thesis. University of Sheffield, UK.

Hadi, A. S. (1992). Identifying multiple outliers in multivariate datasets. J. R. Statistical Society, B, **54**, 761-771.

Johnson, R. A., & Wichern, D. W. (2002). *Applied Multivariate Statistical Analysis*, Prentice Hall, New Jersey.

Nkansah, B. K., & Gordor, B. K. (2012a). A Procedure for Detecting a Pair of Outliers in Multivariate Datasets. *Studies in Mathematical Sciences*, **4**(2), 1 – 9.

Nkansah, B. K., & Gordor, B. K. (2012b). On the One-Outliers Displaying Component. *Journal of Informatics and Mathematical Sciences*, **4**(2), 229 – 239.

Nkansah, B. K., & Gordor, B. K. (2013). Discordancy in Reduced Dimensions of Outliers in High-dimensional Datasets: Application of Updating Formula. *American Journal of Theoretical and Applied Statistics*, **2**(2), 29 – 37.