J. nat. Rubb. Res., 3(1), 54–59

Identifying Anomalous Laboratories in Interlaboratory Crosscheck Programme by Multivariate Outlier Analysis

LEONG YIT SAN*

Anomalous laboratories can be identified by multivariate outlier analysis involving Mahalanobis distance measure. The power in identifying an outlier is increased when higher dimensional multivariate analysis is used resulting in an improvement over Youden's two-sample diagram. On the other hand, the size of the samples and the complexity of the analysis are increased. An outlier tends to lie far out from the main body of points in a graphical display of the first two principal components.

Quite frequently during the process of collecting data from the field, factory or laboratory, the data are contaminated with unrepresentative, rogue or outlying observations. This contamination of data often reduces and distorts the information provided by the data about the source or generating mechanism.

Outliers are described in the simplest form as follows. In a moderate-size sample taken from a certain population, one or two values are surprisingly far away from the main group. It appears to be inconsistent with the remainder of that set of data. The researcher is tempted to throw away the apparently erroneous values even though he is not certain that the values are spurious. On the contrary, there is a positive although extremely small probability that such values will occur in an experiment. The researcher feels that the loss in the accuracy of the experiment caused by throwing away a couple of good values is small compared to the loss caused by keeping even one bad value. His problem, then, is to introduce some degree of objectivity into the rejection of outlying observations.

Statistical procedures or discordancy tests to identify outliers in univariate samples are well documented in statistical literature and books[1]. On the other hand, multivariate outlier analysis is not commonly used by researchers. This paper evaluates the usefulness of a multivariate outlier analysis.

MATERIALS AND METHODS

The Data

The data were obtained from the SMR round robin interlaboratory crosscheck programme. Two types of materials *A* and *B*, were specially prepared for homogeneity. Two preparations were made for each material. Five replicates for each preparation and type of material were sent to each participating laboratory. For each round of testing, the laboratories were further divided into two groups. An outlier can then be viewed as an anomalous laboratory producing inconsistent test results far away from the remaining laboratories.

Use of Mahalanobis Distance Measure

Anomalous laboratories can be identified utilising a generalised distance procedure to screen multivariate data for outliers[1-5]. The procedure can be performed by computing the Mahalanobis distance of each laboratory from the centre of the distribution of the remaining laboratories under the usual assumption of

---

*Rubber Research Institute of Malaysia, P.O. Box 10150, 50908 Kuala Lumpur, Malaysia

homogeneity of variances, independence of errors and normality. If the probability of the F statistic corresponding to the greatest distance is smaller than a specified value (usually 5%), the laboratory involved is removed from the analysis and the process is repeated until all probabilities are sufficiently larger than the specified value.

Computational Procedure

Let $\boldsymbol{x}_i$ denote the m element vector of the ith laboratory. The computational procedure is performed as follows:

Step 1

The following are computed:

Means $$\bar{\boldsymbol{x}} = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{x}_i$$

Crossproduct matrix

$$\boldsymbol{S} = \sum_{i=1}^{n} (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' (\boldsymbol{x}_i - \bar{\boldsymbol{x}})$$

Crossproduct matrix $$\boldsymbol{R} = \boldsymbol{D}\,\boldsymbol{S}\,\boldsymbol{D}$$

where $\boldsymbol{D}$ is a diagonal matrix with elements $d_{jj} = 1/\sqrt{s_{jj}}$

Standard deviations $$s_j = \sqrt{s_{jj}/(n-1)}$$

Step 2

The eigen values β_j, and the eigen vectors v_j of $\boldsymbol{R}$ are computed by the Jacobi method. The first two vectors are used to display the data and the percentage of dispersion attributed to them is calculated.

Step 3

$\boldsymbol{S}$ is inverted by pivoting on diagonal elements.

$$\text{Let } \boldsymbol{A} = \boldsymbol{S}^{-1}$$

Step 4

Starting values for the quantities d_i used in *Steps 6* and *7* are computed:

$$d_i = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})' \boldsymbol{A} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \quad i = 1, \ldots, n$$

Step 5

The tolerance T is defined as the smallest of the values $1/a_{jj}s_{jj}$ $j = 1, \ldots, m$. If T is less than 0.00001, the covariance matrix is assumed to be singular and the process is terminated by going to *Step 8*.

Step 6

The statistic d_i' defined as

$$d_i' = \frac{n\,(n-2)}{(n-1)} \; \frac{d_i}{1 - \frac{1}{n} - d_i}$$

is algebraically equivalent to Mahalanobis D square $(\boldsymbol{x}_i - \boldsymbol{v})' \boldsymbol{C}^{-1} (\boldsymbol{x}_i - \boldsymbol{v})$

Where $\boldsymbol{v}$ and $\boldsymbol{C}$ are the mean vector and covariance matrix including all the laboratories except the ith laboratory and those which have already been removed.

As noted by Gnanadesikan and Kettenring[2] this outlier procedure falling into the class of generalised distance is particularly useful for uncovering laboratories which lie far afield from the general scatter of laboratory points. Since d_i' is an increasing function of d_i, the laboratory with the largest D square after having been removed is the one with the largest d_i'. Let it be denoted by laboratory k. Under the null hypothesis that laboratory k is from the same multivariate normal population as the remaining laboratories, the statistic

$$F = \frac{n-m-1}{m} \; \frac{d_k}{1 - \frac{1}{n} - d_k}$$

has a F distribution with m and n-m-1 degrees of freedom. The probability P, of a F this large or larger is computed. If P is greater than P_c, go to *Step 8*.

Step 7

The vectors $\boldsymbol{h}$ and $\boldsymbol{g}$ are computed:

$$\boldsymbol{h} = (\boldsymbol{x}_k - \bar{\boldsymbol{x}})/(n-1)$$
$$\boldsymbol{g} = \boldsymbol{A}\,\boldsymbol{h}$$

A is replaced by

$$A + \frac{gg'}{c} \quad \text{where } c = \frac{1}{n(n-1)} - h'g$$

For each laboratory i which has not been removed, d_i is replaced by

$$d_i + 2e_i + f + \frac{(e_i + f)^2}{c}$$

where $e_i = (x_i - \bar{x})'g;\ f = h'g$

n is replaced by n-1, $\bar{x}$ is replaced by $\bar{x} - h$

Go to *Step 5*

Step 8

The means and standard deviations of the remaining laboratories are computed and printed.

Step 9

For each laboratory, the values of the first two principal components are plotted with the laboratories which have been removed and identified as follows:

$$C_{ki} = v'\ (x_i - \bar{x})$$

$$k = 1,2;\ i = 1,\ldots,n$$

RESULTS AND DISCUSSION

Table 1 shows the anomalous or outlying laboratories in initial Wallace Plasticity *(P_0)* from Round Robins 37 to 34 using two, three and four dimensional multivariate analysis as described earlier. For each round, preparations A1 and B1 were used in the two dimensional multivariate analysis whereas preparations A1, A2 and B1 were used in the three dimensional multivariate analysis. All four preparations were used in the four dimensional multivariate analysis.

In Round Robin 37-1, three laboratories were classified as anomalous using four dimensional multivariate analysis, while two laboratories using three dimensional multivariate analysis and only one laboratory (17 or F7) using two dimensional multivariate analysis were found to be anomalous. Laboratory Number 17 was found to be anomalous in all three analyses. The probability of the F statistic corresponding to the distance of Laboratory Number 17 from the centre of the remaining laboratories was the smallest when four dimensional multivariate analysis was used and the largest in the case of the two dimensional analysis. In all the other rounds except Rounds 36-2 and 34-2, the four dimensional multivariate analysis produced the smallest probability for an outlying laboratory. This result indicated that the power in identifying the outlying laboratory was increased when higher dimensional multivariate analysis was used. Noting that the sample size was correspondingly increased, this result was not surprising.

A comparison of Table 1 and Table 6.16 of Leong[6] indicated that there were many common anomalous laboratories such as laboratories E7 and C3 in Round 37-1 picked out by both Youden's analysis[7] and the multivariate outlier analysis; however, there were some laboratories not picked out by the former analysis. This difference arose due to small differences in the details of the procedure even though they were similar in mathematical form.

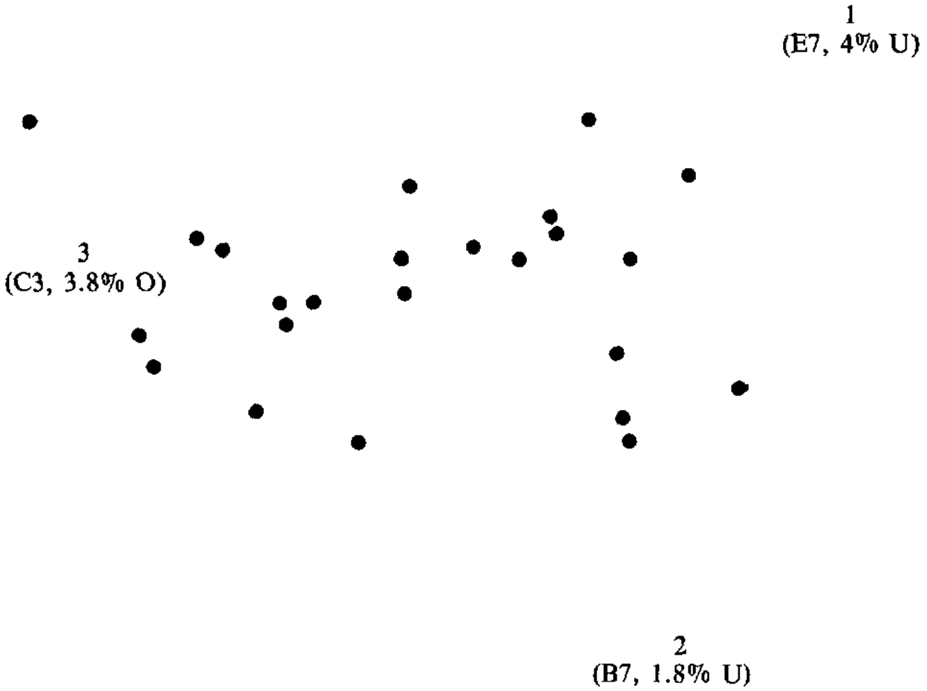
According to Gnanadesikan and Kettenring[2], 'the complexity of the multivariate case suggests that it would be fruitless to search for a truly omnibus outlier protection procedure'. The multivariate case is complex due to the variety of types of multivariate outliers which may arise and the reason that a multivariate outlier can distort not only measures of location and scale but also those of orientation resulting in difficulty in characterising the outlier.

In line with the requirement of displaying the data in graphical form for easy comprehension and interpretation, the first two principal components were used to display the data as shown in *Figure 1* in the four dimensional multivariate case in Round Robin 37-1. It was clear that interpretation of the graph in terms such as tendency to over-estimate or erratic work was not possible. The only observation to be made was that an outlier tended to be lying far out from the main body of points most of the time

TABLE 1. ANOMALOUS LABORATORIES IN INITIAL WALLACE PLASTICITY FROM ROUND ROBINS 37 TO 34 USING TWO, THREE AND FOUR DIMENSIONAL MULTIVARIATE ANALYSIS

Round	Laboratory	Code	Dimension					
			Two		Three		Four	
			F	Prob. (%)	F	Prob. (%)	F	Prob. (%)
37-1	17	E7	3.91	3.54	4.20	1.86	5.43	0.40
	15	B7	—	—	4.63	0.54	4.19	1.34
	14	C3	—	—	—	—	3.24	3.60
37-2	2	C1	—	—	—	—	4.21	1.32
36-1	15	C2	—	—	3.29	4.30	5.65	3.60
	12	C7	—	—	4.00	2.20	3.67	2.36
36-2	16	B3	4.15	3.30	—	—	4.27	1.25
	12	B5	3.64	4.49	5.00	0.95	4.23	1.38
	19	A5	—	—	—	—	3.81	2.19
	10	B1	—	—	—	—	3.51	3.06
	2	E3	3.89	3.66	—	—	—	—
	21	B2	3.94	3.71	—	—	—	—
	18	D6	4.52	2.68	—	—	—	—
35-1	15	BF	—	—	4.57	1.36	6.22	0.22
	18	CG	—	—	3.84	2.75	5.55	0.43
	7	BG	—	—	4.61	1.55	4.29	1.40
	14	BV	—	—	3.62	3.21	3.94	2.06
35-2	22	TC	9.22	0.13	5.90	0.47	6.15	0.24
	3	CI	—	—	—	—	3.18	3.85
34-1	17	CK	—	—	—	—	3.27	3.69
	16	CF	—	—	—	—	5.94	0.40
	6	BD	—	—	—	—	3.55	3.15
34-2	16	BR	3.50	4.99	3.87	2.58	—	—
	19	CB	—	—	4.01	2.38	—	—
	4	BN	3.71	4.38	—	—	—	—
	13	CJ	4.25	3.07	—	—	—	—
	8	BJ	4.02	3.71	—	—	—	—

— Denotes a probability greater than 5%.



Figures within brackets are the laboratory code and percentage over-estimation (O) and under- estimation (U).

Figure 1. Graphical display of laboratories by the first two principal components for Round Robin 37-1.

(the results from other round robins are not presented here). Unlike Youden's approach, the direction of improvement for non-anomalous laboratories could not be indicated from the analysis which was mainly expressed in equation form and the graph of the first two principal components. However, the first two principal components were found to explain nearly all the variations in the five round robins studied as shown in *Table 2.* This means that it was sufficient to concentrate on the first two principal components.

TABLE 2. PERCENTAGE OF DISPERSION ACCOUNTED FOR BY FIRST TWO PRINCIPAL COMPONENTS OF FOUR DIMENSIONAL MULTIVARIATE ANALYSIS IN INITIAL WALLACE PLASTICITY

Round	Percentage dispersion	
	Group 1	Group 2
37	94.13	95.74
36	81.77	81.85
35	95.71	97.76
34	83.18	86.11
23	—	95.36

CONCLUSION

The power in identifying an outlier in the form of an anomalous laboratory in an interlaboratory crosscheck programme is increased when higher dimensional multivariate analysis using the

Mahalanobis distance measure is used. The procedure is an improvement over Youden's two-sample method. On the other hand, the size of the samples and the complexity of the analysis are increased.

Interpretation of the graphical display by the first two principal components cannot be easily made even though they account for nearly all the dispersion in initial Wallace Plasticity for five rounds of crosschecks. The only observation to be made from the display is that an outlier tended to be lying far out from the main body of points most of the time.

REFERENCES

1. BARNETT, V. AND LEWIS, T. (1978) *Outliers in Statistical Data.* John Willey & Sons.
2. GNANADESIKAN, R. AND KETTENRING, J.R. (1972) Robust Estimates, Residuals and Outlier Detection with Multiresponse Data. *Biometrics,* **28**, 81.
3. ROHLF, F.J. (1975) Generalization of the Gap Test for the Detection of Multivariate Outliers. *Biometrics,* **31**, 93.
4. SIOTANI, M. (1959) The Extreme Value of the Generalised Distance of the Individual Points in the Multivariate Normal Sample. *Ann. Inst. Statist. Math., Tokyo,* **10**, 183.
5. WILKS, S.S. (1963) Multivariate Statistical Outlier. *Sankhya Ser. A,* **25**, 407.
6. LEONG, Y.S. (1981) Statistical Methods in International and Local Round Robin Interlaboratory Crosschecks for Standard Malaysian Rubber Testing. Doctor of Agricultural Science Thesis, University of Ghent, Belgium.
7. YOUDEN, W.J. (1959) Graphical Diagnosis of Interlaboratory Test Results. *Industrial Quality Control,* **15(11)**, 24.