MJS, Vol. 7, No. 2, July - Dec 2008, pp. 10-24
ISSN 0975-3311
https://doi.org/10.12725/mjs.13.2

DETECTION OF MULTIDIMENSIONAL OUTLIERS USING BIPLOT ANALYSIS

T.A. Sajesh & M.R. Srinivasan***

ABSTRACT

It is necessary to examine the valuable data being distorted by the presence of outliers before the same is subjected to necessary analysis. Outliers should be identified using reliable detection methods and tested prior to performing data analyses. Detection of outliers in multidimensional data is important in many applications as it will have far reaching consequences in its analysis. There are methods available in the literature for detecting multiple outliers but there exist no unified method for detecting the same. An attempt has been made to detect the multidimensional outliers through Biplot analysis using elliptical method with a well defined axis (a, b) based on Inter Quartile Range (IQR). The performance of the designed methods is examined by a comparison with the existing methods.

1. Introduction

Outliers are observations that appear to be extreme or unusual with respect to the rest of the data and to prior knowledge about what values are plausible. Outliers may be "erroneous" or "real" in the following sense. "Real" outliers are observations whose actual values are, in fact, very different than those observed for the rest of

* *Department of Statistics, University of Madras, Chennai - 600 005.*

** *Department of Statistics, University of Madras, Chennai - 600 005.*

the data and violate plausible relationships among variables. "Erroneous" outliers are observations that are distorted due to misreporting or miss-recording errors in the data-collection process.

Outliers of either type may exert undue influence on the results of statistical analyses, so they should be identified using reliable detection methods prior to performing data analyses. When we encounter a potential outlier, our first suspicion may be that the observation resulted from a mistake or other extraneous effect, and should be discarded. However, if the outlier is "real" rather than "erroneous," it may be conveying important information about the underlying population of real values. Non-judicious removal of observations that appear to be outliers may result in underestimation of the uncertainty present in the data. As a consequence, estimated standard errors and p-values may be smaller than they should be, possibly leading to false findings of significance.

The study of outliers is as important for multivariate data as it is for univariate samples. As Gnanadesikan and Kettenring (1972) remark, a multivariate outlier no longer has a simple manifestation as an observation which 'sticks out at the end' of the sample. A multivariate outlier need not be an extreme in any of its components. Someone who is short and fat need not be the shortest, or the fattest, person around. But that person can still be an 'outlier'. The study of literature reveals that Mahalanobis distance is the basis for the detection of multivariate outliers. The standard method for multivariate outlier detection is robust estimation of the parameters in the Mahalanobis distance and the comparison with a critical value of the Chi square distribution (Rousseeuw and Van Zomeren, 1990). However, also values larger than this critical value are not necessarily outliers; they could still belong to the data distribution. In order to distinguish between extremes of a distribution and outliers, Garrett (1989) introduced the Chi square plot, which draws the empirical distribution function of the robust Mahalanobis distances against the Chi square distribution. A break in the tail of the distributions is an indication for outliers, and values beyond this break are iteratively deleted. The approach of Garrett needs a lot of interaction of the analyst with the data since this method is not an automatic procedure.

A biplot is a graphical display of the rows and columns of a rectangular $n \times p$ data matrix X, where the rows are often the subjects or sample units, and the columns are variables. The concept of Biplots was introduced by Gabriel (1971, 1981) and subsequently developed by Bradu and Gabriel (1978) and Gower (1990, 1992).

In almost all applications, biplot analysis starts with performing some transformation on the data matrix X, depending on the nature of the data. The usual transformations are centering with respect to variable means, with respect to variable medians, normalization of variables etc. The transformed variable Z is being decomposed as $Z = U\Lambda V^T$ using Singular Value Decomposition (SVD).

The concept of SVD is one of the popular decomposition technique used in matrix algebra with wide applications. Under singular value decomposition an n × p matrix A of rank r can be factored into

$$A = U\Lambda V^T$$

Where $\Lambda = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$, with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_r \geq 0$, U is an orthonormal matrix of order n × r, and V an orthonormal matrix of order r × r, i.e. $U^TU = V^TV = I$.

Now, the transformed variable Z be approximated using the first two singular values and corresponding right and left singular vectors

$$Z_2 \approx U_2\Lambda_2V_2^T = GH^T, \text{ where } G = (U_2\Lambda_2^{\alpha}) \text{ and } H = (V_2\Lambda_2^{1-\alpha})$$

where α is a chosen constant with $0 \leq \alpha \leq 1$. The n × 2 matrix H consists of n row vectors representing the rows of the matrix Z. The p × 2 matrix H consists of p column vectors representing the columns of Z. Biplots are usually plotted in two dimensions for ease of display, and hence only the first two singular values and their associated vectors are used in the approximation of Z. Different choices of á give rise to different biplots. The most common choices of á are the values 1 or 0, when the singular values are assigned entirely either to the left singular vectors of U or to the right singular vectors of V respectively, or 0.5 when the square roots of the singular values are split equally between left and right singular vectors. Each choice, while giving exactly the same matrix approximation, will highlight a different aspect of the data matrix. The term "principal coordinates" refers to the singular vectors scaled by the singular values (for example, G with $\alpha = 1$), while "standard coordinates" are the unscaled singular vectors (Greenacre, 1984).

For $\alpha = 1$, that is rows in principal coordinates and columns in standard coordinates, called the *form biplot*, which favours the display of the individuals, while for $\alpha = 0$, that is rows in standard coordinates and columns in principal coordinates, called the *covariance biplot*, which favours the display of the variables (Greenacre and Underhill, 1982). When $\alpha = 0.5$ the biplot favours the display of interaction effect.

In this study biplots have been considered for detecting outliers in multivariate data using elliptical method. Further, a comparative study using different methods (Maximum Likelihood Method and Sequential Method, Barnet and Lewis, 1994), has been considered.

2. Biplot Method

The alternate scheme would be to evolve a method of looking at the distances through an appropriate stretch out of sphere like an ellipse for the detection of outliers. Ellipse would be most adequate as the log-likelihood function of bivariate normal distribution can be viewed as an ellipse (Johnson and Wichern, 1992). Hence, biplots have been considered to provide a solution to the above scheme of drawing an ellipse for identifying outliers.

In biplot method we are using biplot technique for the detection of outliers in multidimensional data. As the first step the data matrix will be transformed by centering with respect to the median vector. Using singular value decomposition we will decompose the transformed matrix in to the product of two matrices and then we will approximate the matrix by the product of two matrices, G and H, of order $n \times 2$ and $2 \times p$ respectively, where n is the number of observations and p is the number of variables. In biplot the G matrix representing the observations and the variables are represented by H matrix. The observations are standardized with the respective standard deviations or Median Absolute Deviations (MAD) before proceeding with biplots.

The detection of outliers is based on the points lying outside an ellipse drawn with parameters (a, b), where a is vertex and b is co-vertex. Mc.Gill et al. (1978) has suggested the use of IQR as a robust measure of spread in data. Hence, the vertices (a, b) based on function of IQR have been identified based on different choices and two methods for detection of outliers have been discussed.

2.1. Mahalanobis Distance - Biplots (MDB) Method

Under this method G matrix is standardized by the corresponding Median Absolute Deviation (MAD) rather than standard deviation. As defined earlier, Mahalanobis distance of standardized G from origin using correlation matrix is considered. However, an ellipse is drawn with a = (1.5/2) * IQR of Mahalanobis Distance and b = (1.25/2) * IQR of Mahalanobis Distances. Biplots are obtained using the standardized G obtained through SVD and if any of the points lie outside the ellipse (a, b) then the presence of outlier is suspected.

2.2. Inter Quartile Range – Biplots (IQRB) Method

In IQRB method G matrix is standardized by the corresponding Median Absolute Deviation (MAD). Then Inter Quartile Range (IQR) of each of the components of standardized G matrix are obtained. Now an ellipse is drawn with the set a = 1.5*IQR of first component and b = 1.5*IQR of second component. Now Biplots are obtained using the standardized G obtained through SVD and if any observation lies outside the ellipse (a, b) an outlier is said to be detected.

3. Examples

3.1 Motivating Example: Yields of Grass

Rothamsted Experimental Station is one of the oldest agricultural research centers. One of the 'classical experiments' is Park Grass in which the growth of grasses has been monitored under various fixed treatment regimes for about 150 years. Table 1 presents the yields of grass (in t/ha dry matter) on two totally untreated plots for the 50 years from 1941 to 1990.

Table 1

Year	plot-3	plot-12	Year	plot-3	plot-12
1941	0.85	1.26	1966	1.43	2.16
1942	0.26	0.59	1967	1.31	1.48
1943	1.03	1.66	1968	1.52	1.28
1944	0.34	0.65	1969	0.72	1.87
1945	1.14	1.75	1970	1.15	1.51
1946	1.18	0.8	1971	1.5	2.94
1947	1.52	1.67	1972	1.4	1.54
1948	1.12	1.25	1973	1.24	1.27
1949	0.62	0.78	1974	1.18	1.25
1950	0.89	0.76	1975	0.91	0.55
1951	1	1.42	1976	1.06	1.22
1952	1.58	1.8	1977	1.2	1.21
1953	1.63	1.84	1978	1.7	1.77
1954	0.99	1.05	1979	1.26	2.27

Year	plot-3	plot-12	Year	plot-3	plot-12
1955	1.1	1.58	1980	0.85	1.07
1956	0.69	1.11	1981	1.47	1.95
1957	0.6	1.02	1982	2.03	1.91
1958	1.21	1.61	1983	0.99	0.84
1959	0.51	0.62	1984	1.08	1.22
1960	1.56	1.82	1985	1.58	1.65
1961	1.39	1.82	1986	0.78	1.02
1962	1.2	1.28	1987	1.39	1.71
1963	1.43	1.64	1988	1.4	1.38
1964	1.48	2.47	1989	0.6	0.75
1965	2.75	3.45	1990	0.88	0.94

Figure 1 is the scatter diagram of these yields. Several observations appear as outliers (notably those for 1965, 1971, 1969, 1982 and 1942) marked as A, B, C, D and E on the figure.

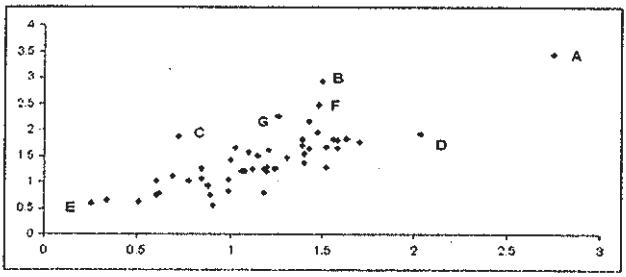

Figure 1

Maximum likelihood (ML) method detected A, B, C, D and E as extreme observations. But the Sequential method suspects the observations A, B, C, F and G as outliers. This method does not project D and E as outliers as suspected by the ML method. On the other hand the suspected outliers F and G identified through the sequential method could not be identified through the ML method. Hence, these methods have not provided an unanimous result on the identification of outliers.

The outputs obtained through the Biplot methods-MDB method and IQRB method-re given below.

a. MDB Method: The MDB method gives the following plot

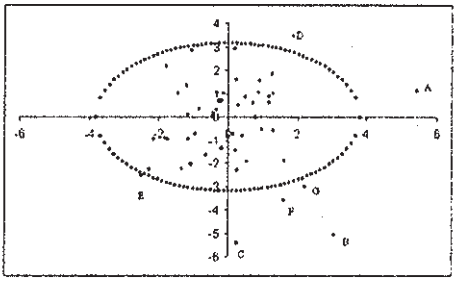

Figure 2

From Figure 2 it can be seen that the points A, B, C, D, F and G are lying outside the ellipse and the observation E is lying on the boundary. So we can suspect A, B, C, D, F and G as outliers. E can also be considered as an extreme observation as it is lying on the boundary.

b. IQRB Method: Figure 3 is the resultant plot obtained from IQRB method.

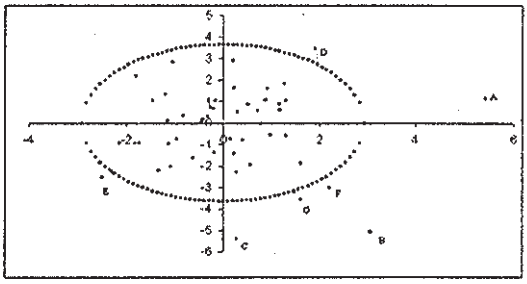

Figure 3

From the above figure it can be seen that the points A, B, C, D, E, F and G are lying outside the ellipse and can be suspected as outliers.

The ML method has shown that A, B, C, D, E as the outliers and the Sequential method has shown that besides A, B, C two additional points namely, F and G are observed as outliers, but D and E are not detected as outliers. However, both the Biplot methods: MDB and IQRB have detected all the seven observations as outliers.

3.2 Example 2

Example 1 discussed above was a 50 × 2 matrix and hence the methods needs to be examined for a larger data set. The data from **Data 1** is a 55 × 7 matrix which presents woman track data set for fifty five countries (Appendix I). The four methods discussed so far have been used for the data set to examine the presence of outliers.

According to ML method the observations Wsamoa, Mauritius, Cookis, Guinea and Guatemal are suspected outliers. Sequential method also identified the observations Wsamoa, Mauritius, Cookis, Guinea and Guatemal as extremities.

The outputs obtained through the Biplot methods are given below.

a. MDB Method: The MDB method gives the following graph.

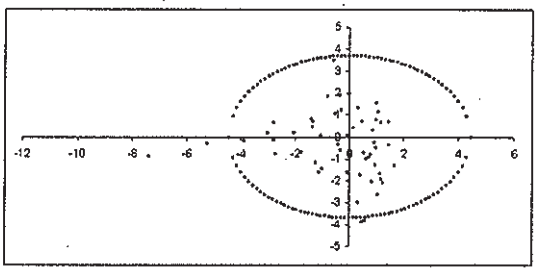

Figure 4

From figure 4 it can be seen that the observations Wsamoa, Mauritius, Cookis, Guinea, Guatemal, GDR, Czech, Costa and Korea are lying outside the ellipse and thus can be suspected as outliers.

b. IQRB Method: Figure 5 shows the resultant plot obtained from IQRB method.

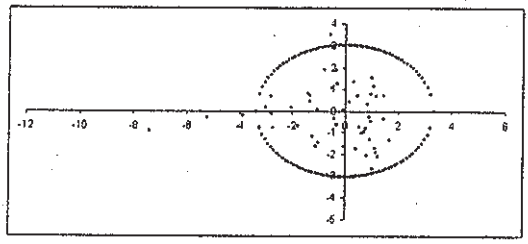

Figure 5

The above graph shows that the observations Wsamoa, Mauritius, Cookis, Guinea, Guatemal, GDR, Czech, Costa and Korea are lying outside the ellipse and thus can be suspected them as outliers.

Both the Biplot methods have detected all the extreme observations obtained from the Maximum likelihood and Sequential method. In addition, the Biplots methods also identified some other observations (GDR, Czech, Costa and Korea) as suspected outliers.

4. Conclusions

The problem of identification of outliers and testing for the same is extremely important in any data set as the presence of it could affect the inference. The detection of outliers in a multidimensional data is fairly complex because of the components involved in it. There are methods like Maximum Likelihood and Sequential methods available for detection of outliers in multivariate data which invariably leads to different results. Hence an attempt has been made to use the distances obtained through biplots and an appropriate stretch of sphere like an ellipse with a well defined axis using IQR for the detection of outliers. The methods designed for detection of outliers are compared with some of the methods available in the literature.

The four methods are illustrated with examples and graphical presentation of outliers. MATLAB coding for IQRB method and MDB method is presented in Appendix II. There is scope for extending the method of detection of outliers through biplots as it based on Singular Value Decomposition. The robustness needs to be examined in detail.

Acknowledgement

The authors would like to thank Naval Research Board, DRDO, Ministry of Defence, New Delhi for the project grant provided in carrying out the research work.

References

1. Aitchison, J., and Greenacre, M. (2002). Biplots for compositional data. Journal of the Royal Statistical Society, Series C (Applied Statistics), 51(4), 375–392.

2. Barnett, V., and Lewis, T. (1994). Outliers in Statistical Data, John Wiley & Sons, Chichester, England.

3. Bradu, D. and Gabriel, K. R. (1978). The biplot as a diagnostic tool for models of two-way tables, Technometrics, 20, 47-68.

4. Gabriel, K. R. (1971). The biplot-graphic display of matrices with application to principal component analysis. Biometrika, 58, 453-467.

5. Gabriel, K. R. (1981). Biplot display of multivariate matrices for inspection of data and diagnosis. Wiley, New York.

6. Gilbert Strang. (2003). Introduction to Linear Algebra, Third Edition. Wellesley-Cambridge Press, USA.

7. Gnanadesikan, R., and Keetenring, J. R. (1972). Robust estimates, residuals and outlier detection with multiresponse data. Biometrics, 28, 81-124.

8. Greenacre, M.J. (1984). Theory and Application of Correspondence Analysis. Academic Press, London.

9. Greenacre, M.J. and Underhill, L.G. (1982). Scaling a data matrix in low-dimensional Euclidean space. Cambridge University Press, Cambridge, UK.

10. Johnson, R. A., and Wichern, D, W. (1992). Applied Multivariate Analysis. Prentice-Hall of India Private Limited, New Delhi.

11. McGill, R., J.W. Tukey, and W. Larsen. (1978). Variations of boxplots. The American Statistician, 32.1, 12-16.

12. Rousseeuw, P.J., and Van Zomeren, B. c. (1990). Unmasking multivariate outliers and leverage points. J. Amer. Statist. Assoc., 85, 633-639.

Appendix I

Data 1

Country	100	200	400	800	1500	3000	Marathon
Argentina	11.61	22.94	54.50	2.15	4.43	9.79	178.52
Australia	11.20	22.35	51.08	1.98	4.13	9.08	152.37
Austria	11.43	23.09	50.62	1.99	4.22	9.34	159.37
Belgium	11.41	23.04	52.00	2.00	4.14	8.88	157.85
Bermuda	11.46	23.05	53.30	2.16	4.58	9.81	169.98
Brazil	11.31	23.17	52.80	2.10	4.49	9.77	168.75
Burma	12.14	24.47	55.00	2.18	4.45	9.51	191.02
Canada	11.00	22.25	50.06	2.00	4.06	8.81	149.45
Chile	12.00	24.52	54.90	2.05	4.23	9.37	171.38
China	11.95	24.41	54.97	2.08	4.33	9.31	168.48
Columbia	11.60	24.00	53.26	2.11	4.35	9.46	165.42
Cookis	12.90	27.10	60.40	2.30	4.84	11.10	233.22
Costa	11.96	24.60	58.25	2.21	4.68	10.43	171.80
Czech	11.09	21.97	47.99	1.89	4.14	8.92	158.85
Denmark	11.42	23.52	53.60	2.03	4.18	8.71	151.75
Dominican	11.79	24.05	56.05	2.24	4.74	9.89	203.88
Finland	11.13	22.39	50.14	2.03	4.10	8.92	154.23
France	11.15	22.59	51.73	2.00	4.14	8.98	155.27
GDR	10.81	21.71	48.16	1.93	3.96	8.75	157.68
FRG	11.01	22.39	49.75	1.95	4.03	8.59	148.53
GB&NI	11.00	22.13	50.46	1.98	4.03	8.62	149.72
Greece	11.79	24.08	54.93	2.07	4.35	9.87	182.20
Guatemal	11.84	24.54	56.09	2.28	4.86	10.54	215.08
Hungary	11.45	23.06	51.50	2.01	4.14	8.98	156.37
India	11.95	24.28	53.60	2.10	4.32	9.98	188.03

Country	100	200	400	800	1500	3000	Marathon
Indonesia	11.85	24.24	55.34	2.22	4.61	10.02	201.28
Ireland	11.43	23.51	53.24	2.05	4.11	8.89	149.38
Israel	11.45	23.57	54.90	2.10	4.25	9.37	160.48
Italy	11.29	23.00	52.01	1.96	3.98	8.63	151.82
Japan	11.73	24.00	53.73	2.09	4.35	9.20	150.50
Kenya	11.73	23.88	52.70	2.00	4.15	9.20	181.05
Korea	11.96	24.49	55.70	2.15	4.42	9.62	164.65
DPRKorea	12.25	25.78	51.20	1.97	4.25	9.35	179.17
Luxembou	12.03	24.96	56.10	2.07	4.38	9.64	174.68
Malásiya	12.23	24.21	55.09	2.19	4.69	10.46	182.17
Mauritius	11.76	25.08	58.10	2.27	4.79	10.90	261.13
Mexico	11.89	23.62	53.76	2.04	4.25	9.59	158.53
Netherlands	11.25	22.81	52.38	1.99	4.06	9.01	152.48
NZealand	11.55	23.13	51.60	2.02	4.18	8.76	145.48
Norway	11.58	23.31	53.12	2.03	4.01	8.53	145.48
Guinea	12.25	25.07	56.96	2.24	4.84	10.69	233.00
Philippi	11.76	23.54	54.60	2.19	4.60	10.16	200.37
Poland	11.13	22.21	49.29	1.95	3.99	8.97	160.82
Portugal	11.81	24.22	54.30	2.09	4.16	8.84	151.20
Rumania	11.44	23.46	51.20	1.92	3.96	8.53	165.45
Singapore	12.30	25.00	55.08	2.12	4.52	9.94	182.77
Spain	11.80	23.98	53.59	2.05	4.14	9.02	162.60
Sweden	11.16	22.82	51.79	2.02	4.12	8.84	154.48
Switzerl	11.45	23.31	53.11	2.02	4.07	8.77	153.42
Taipei	11.22	22.62	52.50	2.10	4.38	9.63	177.87
Thailand	11.75	24.46	55.80	2.20	4.72	10.28	168.45
Turkey	11.98	24.44	56.45	2.15	4.37	9.38	201.08
USA	10.79	21.83	50.62	1.96	3.95	8.50	142.72
USSR	11.06	22.19	49.19	1.89	3.87	8.45	151.22
WSamoa	12.74	25.85	58.73	2.33	5.81	13.04	306.00

Appendix II

% MATLAB Coding for IQRB method

```
m=median(y);
for i=1:size(y,1)
   for j=1:size(y,2)
      x(i,j)=y(i,j)-m(1,j);
   end
end
[u,s,v]=svd(x);
u2=[u(:,1),u(:,2)];
m1=median(abs(u2(:,1)-median(u2(:,1))));
m2=median(abs(u2(:,2)-median(u2(:,2))));
z=[u2(:,1)/m1,u2(:,2)/m2];
z
q11=prctile(z(:,1),25);
q13=prctile(z(:,1),75);
q21=prctile(z(:,2),25);
q23=prctile(z(:,2),75);
iqr1=q13-q11;
iqr2=q23-q21;
a=1.5*iqr1;
```

```
b=1.5*iqr2;
d1=(-a:(a/10):a);
t1=[d1';d1'];
for i=1:size(d1,2)
  d2(i)=b*sqrt(1-((d1(i)^2)/(a^2)));
end
t2=[d2';-d2'];
e=[t1,t2];
graph=[z;e];
scatter(graph(:,1),graph(:,2))
```

% MATLAB Coding for MDB method

```
n=median(y);
for i=1:size(y,1)
  for j=1:size(y,2)
    x1(i,j)=y(i,j)-n(1,j);
  end
end
[k,w,r]=svd(x1);
k2=[k(:,1),k(:,2)];
n1=median(abs(k2(:,1)-median(k2(:,1))));
n2=median(abs(k2(:,2)-median(k2(:,2))));
```

```
z1=[k2(:,1)/n1,k2(:,2)/n2];
z1
ic=inv(corr(z1));
g=z1*ic*z1';
for i=1:size(z1,1)
    md(i)=g(i,i);
end
iqr=prctile(md',75)-prctile(md',25);
aa=1.5*iqr/2;
bb=1.25*iqr/2;
d11=(-aa:(aa/10):aa);
t11=[d11';d11'];
for i=1:size(d11,2)
   d22(i)=bb*sqrt(1-((d11(i)^2)/(aa^2)));
end
t22=[d22';-d22'];
e1=[t11,t22];
graph1=[z1;e1];
scatter(graph1(:,1),graph1(:,2))
```