Sains Malaysiana 43(12)(2014): 1973–1977

Eigenstructure-Based Angle for Detecting Outliers in Multivariate Data

(Sudut Berasaskan Struktur Eigen untuk Mengesan Titik Terpencil dalam Data Multivariat)

NAZRINA AZIZ*

ABSTRACT

There are two main reasons that motivate people to detect outliers; the first is the researchers' intention; see the example of Mr Haldum's cases in Barnett and Lewis. The second is the effect of outliers on analyses. This article does not differentiate between the various justifications for outlier detection. The aim was to advise the analyst about observations that are isolated from the other observations in the data set. In this article, we introduce the eigenstructure based angle for outlier detection. This method is simple and effective in dealing with masking and swamping problems. The method proposed is illustrated and compared with Mahalanobis distance by using several data sets.

Keywords: Angle; Eigenstructure; masking; outliers; swamping

ABSTRAK

Terdapat dua sebab utama yang mendorong orang ramai untuk mengesan titik terpencil, yang pertama adalah hasrat penyelidik; lihat contoh kes Encik Haldum di Barnett dan Lewis. Yang kedua adalah kesan titik terpencil ke atas analisis. Kertas ini tidak membezakan antara pelbagai justifikasi untuk mengesan titik terpencil. Tujuannya adalah untuk berkongsi dengan penganalisis mengenai cerapan yang terpencil daripada cerapan lain dalam set data. Dalam kertas ini, kami memperkenalkan sudut berasaskan struktur eigen untuk mengesan titik terpencil. Kaedah ini adalah mudah dan berkesan dalam berurusan dengan masalah litupan dan limpahan. Kaedah yang dicadangkan digambarkan dan dibandingkan dengan jarak Mahalanobis menggunakan beberapa set data.

Kata kunci: Limpahan; litupan; struktur eigen; sudut; titik terpencil

INTRODUCTION

The identification of outliers is very important because of its effect to the analysis finding. If the statistical models are simply applied to the data sets containing outliers, one might get a misleading result. For example, in the regression analysis, one of the effects of the appearance of outliers is that they would control the regression line where the outliers will pull the regression line in their direction. In other words, it will influence the regression coefficient, which might calculate all the predicted values to wrong values. Many authors have discussed these issues critically (Chatterjee & Hadi 1988; Cook & Weisberg 1982; Rousseeuw & Leroy 1987).

In the case of principle component analysis or factor analysis, the existence of outliers will deflate the correlation coefficient and this will automatically influence the factor score (Wulder 2002). The similar problem can also happen to an analysis of variance; the appearance of outliers might prove a large influence on the estimate of variance and this can cause a low probability of rejecting the hypothesis since it will affect the F statistics value (Quinn & Keough 2002). Outliers are also known as a special target of interest in the realistic environment. Hodge (2004) listed a few applications that implemented outlier detection. For example, in the monitoring activity, one can detect mobile phone deception by monitoring phone activity or suspicious trades in the equity market, while in the loan application processing, one can identify a potentially problematic customer.

There have been many methods developed for the identification of outliers. They can be classified into the univariate method and the multivariate method (Barnett & Lewis 1994; Hawkins 1980). The univariate method is performed independently on each variable, whereas the multivariate method investigates the relationship of several variables (Franklin et al. 2000). One cannot claim multivariable observations as outliers if each variable is considered independently. This makes the identification of outliers become more difficult in the higher dimension data.

Some of the multivariate outlier detection methods have been modified from the univariate method, so that it can take into account a multivariable. Examples are the generalized distance with studentized residual (Siotani 1959), the ratio of generalized distance with all observations (Wilk 1963) and the W statistics for normality (Shapiro & Wilk 1965). Wilks statistics (Wilk 1963) is also widely used for identification of outliers. It is equivalent to using the Mahalanobis distance of the n sample points, from the sample mean (Caroni & Billor 2007). However, this method is subject both to the masking and swamping effect when a data set contains clustered outliers.

The masking problem occurs when the appearance of one outlier covers the appearance of another outlier, whereas the swamping problem arises when the observation is identified as an outlier even if it is not (Hawkins et al. 1984). This consideration makes it desirable to consider a robust method of identifying outliers such as minimum volume ellipsoide (MVE) estimators (Rousseeuw & von Zomeren 1990) and minimum covariance determinant (MCD) estimators by Rousseeuw and Driessen (1999).

Robust estimators have the desirable properties of high breakdown point and affine equivariant. The breakdown point is a percentage of outliers that can cause an estimator to take arbitrary large values (Hampel 1971). Therefore, estimators with a large breakdown point are more robust. Another desirable property of an estimator is affine equivariant. If an estimator is affine equivariant, stretching or rotating the data will not affect the estimator. Nevertheless, it is noted that the multivariate robust measures suffer from computational complexity, i.e. the efficiency of algorithms as run time and memory requirement permit.

Alternatively to robust approach, this study proposed a method for identification of outliers using eigenstructure based angle. The idea of using the eigenstructure based angle as a tool for identification of outliers is motivated by maximum eigen difference (MED). Given that

$$\mathrm{MED}_i = \frac{d'_j}{\sum_{j=1}^{n} d'_j},$$

where $d'_i = \left\| \lambda_1^{(i)} v_1^{(i)} - \lambda_1 v_1 \right\| \left(1 - \prod_{k=1}^{p} I'_{\left(y_{ik}^2 < \lambda_k\right)} \right)$ and $\| \cdot \|$ represent the euclidean norm. $I'_{\{\cdot\}}$ is an indicator function and $y_{ik} = (x_i - \bar{x})^T v_k$. $\lambda^{(i)}$ and $v^{(i)}$ is an eigenvalues and eigenvectors, respectively, calculated from covariance matrix of data set, X with p dimensions where the ith observation has been removed from it.

The function of $1 - \prod_{k=1}^{p} I'_{\left(y_{ik}^2 < \lambda_k\right)}$ is to let MED_i become zero if all y_{ik}^2 is less than corresponding λ_k where $k = 1, 2, \ldots, p$. This is because if x_is are close to mean, $\bar{x}$ they should not be identified as outliers and their proportion with $y_{ik}^2 < \lambda_k$ for all k is not large if all observations x_i are identically and independently distributed with normal distribution (Goa et al. 2005).

This method utilizes the maximum eigenvalue and the corresponding eigenvector. It is noted that examination of the observations effect on the maximum eigenvalue is very significant. The reason is that outliers that lie in the direction close to the maximum eigenvalue or vice versa, will change the maximum eigenvalue (Goa et al. 2005). The maximum eigenvalue contains maximum variance, therefore, the outliers detected by the maximum eigenvalue have a greater effect on variance and they need extra attention.

The main objective of this paper was to introduce the eigenstructure based angle for detecting outliers. The method is formulated in the next section. In the section that follows, some illustrative examples are given before we conclude.

THE ANGLE

Let $\boldsymbol{X}^T\boldsymbol{X}$ have the eigenvalues-eigenvectors pair (λ_1, v_1), (λ_2, v_2), …, (λ_p, v_p), where $\mathbf{X}$ is an $n \times p$ observation matrix consisting of n observations for p variables. If ith row of matrix $\boldsymbol{X}$ is deleted, one can write it as $\boldsymbol{X}_{(i)}$ where the subscript i in parentheses is read as 'with observation i is removed from $\boldsymbol{X}$', i.e. the ith row of $\boldsymbol{X}$ is x_i^T then $\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)} = \boldsymbol{X}^T\boldsymbol{X} - x_i x_i^T$. Let $\boldsymbol{X}_{(i)}^T\boldsymbol{X}_{(i)}$ have the eigenvalues and eigenvectors pair $(\lambda_{1(i)}, v_{1(i)})$, $(\lambda_{2(i)}, v_{2(i)})$, …, $(\lambda_{p(i)}, v_{p(i)})$. Now, consider the relationship between eigenstructure as follows:

The relationship of eigenvalues λ_j and $\lambda_{j(i)}$ is given by

$$\lambda_{j(i)} = \frac{1}{n-1}\left(l_{ij}^2 - \lambda_j\right) - \frac{1}{2\left(n-1\right)^2} l_{ij}^2 \left[1 + \sum_{k \neq j} \frac{l_{ij}^2}{\lambda_k - \lambda_j} \right] + O\left(\frac{1}{n^3}\right),$$

where $l_{ij} = (x_i - \bar{x})^T vj;$

The relationship between eigenvectors of v_j and $v_{j(i)}$ is obtained based on the observation matrix $\boldsymbol{X}$ given by Goa et al. (2005) as follows:

$$v_{j(i)} = v_j + \frac{l_{ij}}{n-1} \sum_{k \neq j} \frac{l_{ik} v_k}{\lambda_k - \lambda_j} - \frac{1}{2\left(n-1\right)^2} \sum_{k \neq j} \left[\frac{l_{ij}^2 l_{ik}^2 v_j}{\left(\lambda_k - \lambda_j\right)^2} - \frac{2 l_{ik}^2 l_{ij}}{\left(\lambda_k - \lambda_j\right)} \sum_{k \neq 1} \frac{l_{ik} v_k}{\lambda_k - \lambda_j} + \frac{2 l_{ij}^3 v_k}{\left(\lambda_k - \lambda_j\right)^2} \right] + O\left(\frac{1}{n^3}\right).$$

One can develop the angle between v_j and $v_{j(i)}$ (Mertens 1998). If th is an outlier, therefore v_j will change when ith observation is deleted from the sample data matrix, $\mathbf{X}$. Let $\theta_{j(i)}$ be the angle between the jth eigenvectors of S for the given data $\boldsymbol{X}$ and the $j(i)$th eigenvectors when the ith observation is deleted in $\boldsymbol{X}$ (i.e., $\boldsymbol{X}_{(i)}$), then one has the formulae of $\theta_{j(i)}$ by Wang and Nyquist (1991) as $\cos(\theta_{j(i)}) = \frac{1}{2}\left\| v_j + v_{j(i)} \right\|^2 - 1$, or it can be re-written as a function of eigenvalues and eigenvectors by $\theta_{j(i)} = \cos^{-1}\left\{ \frac{l_{ij}/\lambda^*_{j(i)}}{\sqrt{\sum_{k=1} l_{ik}^2 / \left(\lambda^*_{j(i)} + \left(\lambda_k - \lambda_j\right)\right)^2}} \right\}$, where $j = 1, 2, \ldots, p$; $i = 1, 2, \ldots, n$. l_{ij} is the principal component scores of the omitted observation in the principal component decomposition of the complete data $\boldsymbol{X}$ and

$$\lambda^*_{j(i)} = \lambda_j - \frac{1}{n-1}\left(l_{ij}^2 - \lambda_j\right) - \frac{1}{2\left(n-1\right)^2} l_{ij}^2 \left[1 + \sum_{k \neq 1} \frac{l_{ij}^2}{\lambda_k - \lambda_j} \right] + O\left(\frac{1}{n^3}\right).$$

The vector angle is defined as the angle between 0 and 180% that satisfies the relationship $v_j^T v_{j(i)} = \left\|v_j\right\|\left\|v_{j(i)}\right\| \cos\theta_{j(i)}$ where $\left\|.\right\|$ refers to the vector length. If the m observations are deleted from $\boldsymbol{X}$, therefore:

$$\begin{aligned}\theta_{j(I)} &= \cos^{-1}\left\{\frac{v_j^T v_{j(I)}}{\left\|v_j\right\|\left\|v_{j(I)}\right\|}\right\} \\ &= \cos^{-1}\left\{v_j^T\left[v_j + \frac{m}{n-m} l_{jI}\sum_{k\neq j} l_{kI}\left(\lambda_k - \lambda_j\right)^{-1} v_k + \frac{m}{n}\sum_{k=1} v_k^T S_I v_j \left(\lambda_k - \lambda_j\right)^{-1} v_k\right]\right\} \\ &= \cos^{-1}\left\{1 + v_j^T\left[\frac{m}{n-m} l_{jI}\sum_{k\neq j} l_{kI}\left(\lambda_k - \lambda_j\right)^{-1} v_k + \frac{m}{n}\sum_{k=1} v_k^T S_I v_j \left(\lambda_k - \lambda_j\right)^{-1} v_k\right]\right\}.\end{aligned} \tag{1}$$

where $v_j(I) = v_j + \frac{m}{n-m} l_{jI}\sum_{k\neq j} l_{kI}\left(\lambda_k - \lambda_j\right)^{-1} v_k + \frac{m}{n}\sum_{k\neq j} v_k^T S_I v_j(\lambda_k - \lambda_j)^{-1} v_k$ and $l_{jI} = v_j^T\left(\tilde{x}_I - \tilde{x}\right)$ is the mean of principal component score l_{ji_m}, $i_m \in$ I. Note that $v_{j(I)}$, l_{jI} and $l_{ji_m} v_j^T\left(x_{i_m} - \tilde{x}\right)$ are given by Wang and Liski (1993).

Supposing that one only deletes ith observation and considers the maximum eigenvalue, replacing $j = 1$ in (1) leads to

$$\theta_{1(i)} = \cos^{-1}\left\{\frac{l_{i1}/\lambda^*_{1(i)}}{\sqrt{\sum_{k=1} l_{ik}^2 / \left(\lambda^*_{1(i)} + \left(\lambda_k - \lambda_1\right)\right)^2}}\right\}, \tag{2}$$

Next, one can apply the angle, $\theta_{1(i)}$ to identify the outlier in the data set; note that there are a few criteria that will control $\theta_{j(i)}$ value:

First, consider $\lambda_j \geq \lambda_{j(i)}$ and $\lambda_{j(i)} \geq \lambda_{k+1}$ where $j, k = 1, 2, \ldots, p$. One finds that the $\theta_{j(i)}$ value is dominated by the first component of the denominator, i.e. $l_{i1}^2 / \left\{\lambda^*_{1(i)}\right\}^2$. If one substitutes $k = 1$, into $l_{i1}^2 / \left(\lambda^*_{1(i)} + \left(\lambda_k - \lambda_1\right)\right)^2$, hence it becomes $l_{i1}^2 / \left\{\lambda^*_{1(i)}\right\}^2$. Notice that $\frac{l_{ik}^2}{\left\{\lambda^*_{1(i)} + (\lambda_k - \lambda_1)\right\}} > \frac{l_{i(k+1)}^2}{\left\{\lambda^*_{1(i)} + (\lambda_{k+1} - \lambda_2)\right\}}$, and the $\frac{l_{i(k+1)}^2}{\left(\lambda^*_{1(i)} + (\lambda_{k+1} - \lambda_1)\right)}$ value is always small because the denominator is $\{\lambda^*_{1(i)} + (\lambda_{k+1} - \lambda_1)\}$ usually large following $\lambda_{j(i)} \geq \lambda_{k+1}$. As a consequence, if the numerator value of (2) is close to one, the denominator value will also be almost the same; note that the numerator value is always less than the denominator value. This follows that the $\theta_{j(i)}$ yields almost a zero degree angle. Another point is that the value of cos $(\theta_{1(i)})$ is always between -1 and 1.

Next, if the principal component score is negative, $\theta_{1(i)}$ will be large. This corresponds to a negative cosine yielding a large angle.

Therefore, the supposed potential outliers will be situated far away than the remaining observations in the data set if:

$\theta_{1(i)}$ for ith observation is larger than other observations following that $\{\lambda^*_{1(i)}\}$(2) in the first component of ith observation is large; or $\theta_{1(i)}$ for ith observation is smaller than other observations corresponding to $\{\lambda^*_{1(i)}\}$ in the first component of ith observation is small;

The principal component score for ith observation is negative while others are positive. Note that the negative principal component score produces larger $\theta_{1(i)}$ than the positive principal component score and vice versa. Observations in the data set have negative principle component scores, $\theta_{1(i)}$ is larger if ith observation has large cos $(\theta_{1(i)})$.

The outliers can be displayed by the index plot $\{i, \theta_{1(i)}\}$. Based on the angle $\theta_{1(i)}$, the following algorithm is proposed to find outliers:

Find S and $S_{(i)}$; Next find the eigenstructure of S and $S_{(i)}$ and choose the maximum eigenpair (v_1, λ_1) and respectively; Find the principal component score, l_{ik} for each p or compute $l_{ik} = \left(x_i^T v_k\right)$; Compute $\theta_{1(i)}$ and Identify the outlier from the index plot of $\{i, \theta_{1(i)}\}$.

The ith observation is considered as a potential outlier by $\theta_{1(i)}$ if it is located at the top of the index plot $\{i, \theta_{1(i)}\}$.

EXAMPLES

In this section we examine the effectiveness of the angle. We consider three data sets from Rousseeuw and Leroy (1987). First we examined the performance of Mahalanobis distance to the three data sets. Figure 1 contains the index plot of Mahalanobis distance for the three data sets. The solid circle in Figure 1 denotes the observation that supposed to be outlier. As one can see, the Mahalanobis distance fails to detect all outliers known to be present in the three data sets.

Example 1 (Hawkins, Bradu and Kass Data). This artificial data set corresponds to a sample of 75 observations in 3 dimensions. It provides a good example of the masking effect. The index plot for Mahalanobis distance in Figure 1 shows only observation 14 as outlier. It masks all the other outliers. The index plot for angle in Figure 2 manages to unmask all the 14 outliers. The results agree well with Atkinson (1994) Pena and Prieto (2001) and Rocke and Woodruff (1996).

Example 2 (Stack Loss Data). This data set contains 21 observations in 3 dimensions. It is about the operation of a plant for the oxidation of ammonia to nitric acid (Rousseeuw & Leroy 1987). According to Atkinson (1994), Hadi (1992) and Rousseeuw and von Zomeren (1990), observations 1, 2, 3 and 21 are outliers. The index plot for

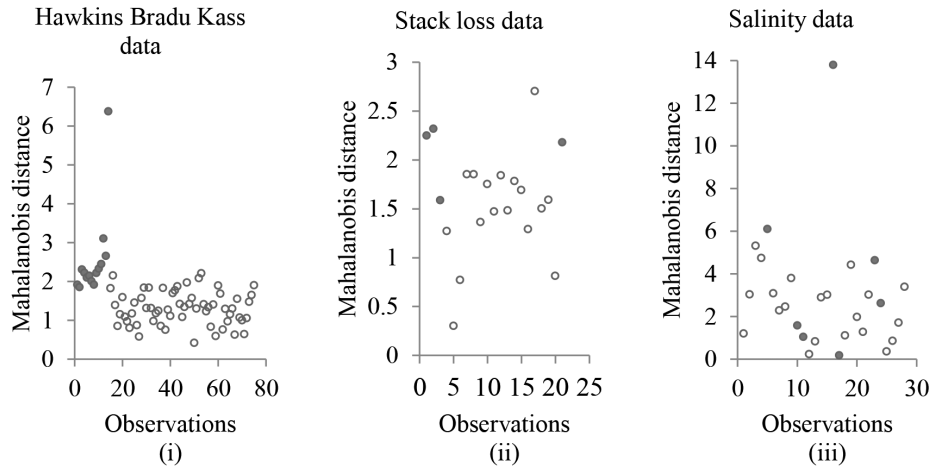

FIGURE 1. Index plots of Mahalanobis distance for (i) Hawkins Bradu Kass data (ii) Stack loss data and (iii) Salinity data

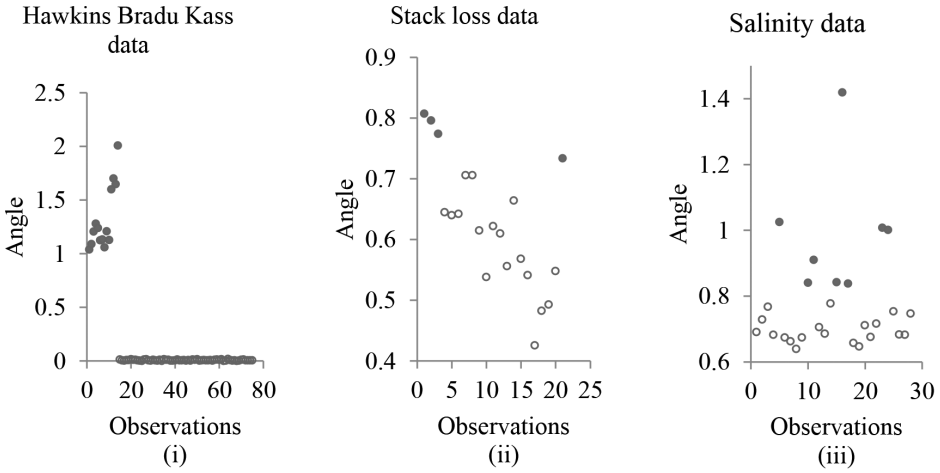

FIGURE 2. Index plots of angle for (i) Hawkins Bradu Kass data (ii) Stack loss data and (iii) Salinity data

Mahalanobis distance (Figure 1) fails to identify any of the many outliers known to be appear in this data set whereas the index plot of angle (Figure 2) claim observations 1, 2, 3 and 21 as outliers.

Example 3 (Salinity Data). The salinity data set contains 28 measurements of water salinity and river discharge taken in North Carolina's Pamlico Sound. Rousseeuw and Leroy (1987) mentioned observations 3, 5 and 16 as outliers in the data set, whereas Pena and Prieto (2001) declares eight observations as the outliers in this data set. The index plot for angle in Figure 2 shows similar finding with Pena and Prieto (2001).

CONCLUSION

In this paper we have proposed eigenstructure based angle for detecting outliers. In the section of examples, we have seen that the Mahalanobis distance is not effective in detecting outliers as it suffers from masking and swamping problems. The eigenstructure based angle manages to identifying outliers. The angle procedure is simple and it can handle the masking and swamping problems.

ACKNOWLEDGEMENTS

We greatly appreciate the helpful comments of the anonymous referees and editor. Their comments have contributed in the improvement of this article. The work that led to the publication of this paper was funded by the Research Grant Scheme of Universiti Utara Malaysia.

REFERENCES

Atkinson, A.C. 1994. Fast very robust methods for the detection of multiple outliers. *Journal of the American Statistical Association* 89(428): 1329-1339.

Barnett, V. & Lewis, T. 1994. *Outliers in Statistical Data*. New York: Wiley and Sons.

Caroni, C. & Billor, N. 2007. Robust detection of multiple outliers in grouped multivariate data. *Journal of Applied Statistics* 34(10): 1241-1250.

Chatterjee, S. & Hadi, A.S. 1988. *Sensitivity Analysis in Linear Regression*. United States: John Wiley.

Cook, R.D. & Weisberg, S. 1982. *Residuals and Influence in Regression*. New York: Chapman and Hall.

Franklin, S., Thomas, S. & Brodeur, M. 2000. Robust multivariate outlier detection using Mahalanobis distance and modified Stahel-Donoho estimators. *Proceeding International Conference on Establishment Surveys*, New York. pp. 697-706.

Gao, S., Li, G. & Wang, D.Q. 2005. A new approach for detecting multivariate outliers. *Communication in Statistics-Theory and Method*. 34: 1857-1865.

Hadi, A.S. 1992. Identyfying multiple outliers in multivariate data. *Journal Royal Statistics Soc. B*. 54(3): 761-777.

Hampel, F.R. 1971. A general qualitative definition of robustness. *Annals of Mathematics Statistic* 42(6): 1887-1896.

Hawkins, D.M. 1980. *Identification of Outliers*. London: Chapman and Hall.

Hawkins, D.M., Bradu, D. & Kass, G.V. 1984. Location of several outliers in multiple regression data using elemental sets. *Technometrics* 26(3): 197-208.

Hodge, V.J. 2004. A survey of outlier detection methodologies. *Artificial Intelligence Review* 22(2): 85-126.

Mertens, B.J.A. 1998. Exact principle component influence measure applied to the analysis of spectroscopic data on rice. *Applied Statistics* 47(4): 527-542.

Pena, D. & Prieto, F.J. 2001. Multivariate outlier detection and robust covariance matrix estimation. *Technometrics* 43(3): 286-299.

Quinn, G.P. & Keough, M.J. 2002. *Experimental Design and Data Analysis for Biologists*. Cambridge: Cambridge University Press.

Rocke, D.M. & Woodruff, D.L. 1996. Identification of outliers in multivariate data. *Journal of the American Statistical Association* 91(435): 1047-1061.

Rousseeuw, P.J. & Driessen, K.V. 1999. A fast algorithm for the minimum covariance determinant estimator. *American Statistical Association and the American Society for Quality* 41(3): 212-223.

Rousseeuw, P.J. & Leroy, A.M. 1987. *Robust Regression and Outlier Detection*. New York: John Wiley.

Rousseeuw, P.J. & von Zomeren, B.C. 1990. Unmasking multivariate outliers and leverage points. *Journal of the American Statistical Association* 85(411): 633-639.

Shapiro, S.S. & Wilk, M.B. 1965. An analysis of variance test for normality (complete samples). *Biometrika* 52: 591-611.

Siotani, M. 1959. The extreme value of the generalized distance of the individual points in the multivariate normal sample. *Annals of the Institute of Statistical Mathematics* 10: 183-208.

Wang, S.G. & Liski, E.P. 1993. Effects of observations on the eigensystem of a sample covariance matrix. *Journal of Statistical Planning and Inference* 36: 215-226.

Wang, S.G. & Nyquist, H. 1991. Effects on the eigenstructure of a data matrix when deleting an observation. *Computational Statistics and Data Analysis* 11(2): 179-188.

Wulder, M. 2002. *A Practical Guide to the Use of Selected Multivariate Statistics*. Victoria: Canadian Forest Service.

Wilk, S.S. 1963. Multivariate statistical outliers. *Sankhya* 25: 407-426.

UUM College of Arts and Sciences
Universiti Utara Malaysia
06010 Sintok, Kedah
Malaysia

*Corresponding author; email: nazrina@uum.edu.my

Received: 20 February 2013
Accepted: 2 May 2014