CSCanada
Studies in Mathematical Sciences
Vol. 4, No. 2, 2012, pp. 1-9
DOI: 10.3968/j.sms.1923845220120402.1623

ISSN 1923-8444 [Print]
ISSN 1923-8452 [Online]
www.cscanada.net
www.cscanada.org

A Procedure for Detecting a Pair of Outliers in Multivariate Dataset

B.K. Nkansah[a],*; B.K. Gordor[a]

[a]Department of Mathematics and Statistics University of Cape Coast, Ghana
*Corresponding author.
Address: Department of Mathematics and Statistics University of Cape Coast, Ghana

Received 22 January, 2012; accepted 19 May, 2012

Abstract
The paper presents a procedure for detecting a pair of outliers in multivariate data. The procedure involves a reduction of the dimensionality of the dataset to only two dimensions along outlier displaying components, and then determines the orientation of a least squares ellipse that fits the scatter of points of the two dimensional dataset. Finally, the reduced data is projected unto a vector which is determined in terms of the orientation of the ellipse. The results show that if two observations constitute a pair of outliers in a data set, then the pair is extreme at either ends of the one-dimensional projection and separated clearly from the remaining observations. If the two outliers are not distinct on such a one-dimensional projection, three key rules are prescribed for successful determination of the right pair of outliers.

Key words
Multiple Outlier Detection; Outlier Displaying Component

B.K. Nkansah, B.K. Gordor (2012). A Procedure for Detecting a Pair of Outliers in Multivariate Dataset. *Studies in Mathematical Sciences, 4*(2), 1-9. Available from: URL: http://www.cscanada.net/index. php/sms/article/view/j.sms.1923845220120402.1623 DOI: http://dx.doi.org/10.3968/j.sms.1923845220120402.1623

1. INTRODUCTION

The detection of a single outlier in a multivariate dataset is one on which several methods converge. One main approach is the use of the Mahalanobis generalised squared distance

$$U(\mathbf{x}_\epsilon, \mathbf{S}_\epsilon) = (\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \tag{1}$$

On the detection of multiple outliers, several procedures have been developed. Most of these methods (e.g. Rousseeuw, 1985; Rousseeuw & Leroy, 1987) rely on the generalised distance in Equation (1) which is often used as a test statistic. Others (e.g. Caroni & Prescott, 1992) make use of the Wilk's ratio statistic which determines the subset T_k of k observations among the sample of size n for which the ratio $R_T = \dfrac{|\,\mathbf{S}_{(T_k)}\,|}{|\,\mathbf{S}\,|}$ is minimum, where $\mathbf{S}$ is the matrix of sum of squares and cross-product and $\mathbf{S}_{(T_k)}$ is the corresponding matrix with the observations in T_k removed from the sample. The results of some of these methods are divergent particularly on the problem of multiple outlier detection. A major cause of this divergence lies in the different approaches adopted at controlling the use of the general sample mean in the detection procedure. Another approach is the one by Gordor and Fieller (1999). Their approach, which is a graphical technique and referred to as the Outlier Displaying Component (ODC), is actually a displaying technique rather than a detecting one. This means that in the case of two outliers, for example, the outliers must be specified before the technique can be used to display them. However, very often these multiple outliers are not known.

We can identify some two basic difficulties that are common to these and many other attempts at multiple outlier detection. These are the amount of computations involved and the subjectivity that characterize these methods.

In the next section we will discuss a procedure for detecting a pair of outliers from a one-dimensional plot of some univariate equivalent of the multivariate dataset. The process, which is a graphical approach, combines the technique of outlier displaying component and minimum ellipse fitting. The technique is illustrated by using some artificial data (see Appendix) which are generated from $N(\mathbf{0},\mathbf{I})$ and consist of 50 four-dimensional observations. The reason for the choice of this artificial dataset will become apparent. In the end, we compare observations that constitute a pair of outliers in some well-known datasets by means of the proposed method and those obtained by the Wilk's ratio statistic.

2. DESCRIPTION OF THE PROCEDURE

The procedure is in three major phases. The first phase broadly deals with dimensionality reduction of the data set from p-dimensional ($p \geq 3$) to 2-dimensional. The second deals with the determination of the orientation of a least squares ellipse that fits the scatter of points of the two dimensional dataset. The last phase then considers a projection vector in terms of the orientation of the ellipse obtained in the second phase. The three phases are expanded in this section.

2.1 Dimensionality Reduction of the Dataset

Suppose that $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n)$ is a p-dimensional random sample with sample mean $\bar{\mathbf{x}}$. Let the single outlier in the sample be $\mathbf{x}_\epsilon$. The One-Outlier Displaying Component (ODC) vector [4] is given by

$$\beta = \mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \tag{2}$$

where $\mathbf{S}$ is $p \times p$ sample sum of squares and cross-product matrix. Now, define the vector $\mathbf{1} = (1, 1, \cdots, 1)'_{n\times 1}$. Then a projection of the mean-corrected data on β gives a univariate equivalent of the data given by

$$\mathbf{t}_1 = (\mathbf{X} - (\mathbf{1} \times \bar{\mathbf{x}}')) \times \beta \tag{3}$$

Now, β is actually an eigenvector of the only non-zero eigenvalue of the matrix $\mathbf{E} = \mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}})(\mathbf{x}_\epsilon - \bar{\mathbf{x}})'$. This non-zero eigenvalue represents the squared generalised distance in Equation (1) of $\mathbf{x}_\epsilon$ from the sample mean, $\bar{\mathbf{x}}$. Now, let $\mathbf{V}_{p\times p} = (\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p)$ be a matrix whose columns are eigenvectors corresponding to the eigenvalues of $\mathbf{E}$, with the first column being β. We apply the Gram-Schmidt Orthogonalization to convert the independent vectors of $\mathbf{V}$ into a set of orthogonal vectors, $\mathbf{w}_k$; $(k = 1, 2, \cdots, p)$ as follows: First, define $\mathbf{w}_1 = \mathbf{v}_1$. Then each $\mathbf{w}_k$ is made orthogonal to the preceding $\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_{k-1}$ by the relation

$$\mathbf{w}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} \frac{\mathbf{v}'_k \mathbf{w}_j}{||\mathbf{w}_j||^2} \mathbf{w}_j, \qquad k = 2, 3, \cdots, p \tag{4}$$

Since $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_p$ are linearly independent, $\mathbf{w}_k \neq 0$; $k = 1, 2, \cdots, p$ and $\mathbf{v}'_k \mathbf{w}_j = 0,\ k \neq j$ Let

$$\mathbf{u}_k = \frac{1}{||\mathbf{w}_k||} \mathbf{w}_k, \qquad k = 1, 2, \cdots, p$$

Thus, $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p)$ are orthonomal. Now the first vector $\mathbf{u}_1$ of $\mathbf{U}$ completely contains all information on the single outlier, $\mathbf{x}_\epsilon$. It will therefore not be useful to use $\mathbf{u}_1$ in the determination of a second component that contains information about the second outlier. Hence, we create a matrix $\mathbf{U} \backslash \mathbf{u}_1 = (\mathbf{u}_2, \mathbf{u}_3, \cdots, \mathbf{u}_p)_{p\times p-1}$. After the orthogonalization, an $(n - 1) \times p$ transformed data set, $\mathbf{Y}_{(\epsilon)} = \mathbf{X} \setminus \mathbf{x}_\epsilon \times \mathbf{U} \setminus \mathbf{u}_1$ is obtained which excludes that of $\mathbf{x}_\epsilon$. Suppose $\bar{\mathbf{y}}$ is the mean vector and $\mathbf{y}_\eta$, the single outlier in $\mathbf{Y}_{(\epsilon)}$ obtained the usual way. Using $\bar{\mathbf{y}}$ and $\mathbf{y}_\eta$, the displaying component for $\mathbf{Y}$ is found as $\varphi = \mathbf{S}_{\mathbf{Y}}^{-1}(\mathbf{y}_\eta - \bar{\mathbf{y}})$, where $\mathbf{S}_{\mathbf{Y}}$ is the sum of squares and cross-product matrix of $\mathbf{Y}_{(\epsilon)}$. The component φ of dimension $p-1$, is referred to as the Sub-Outlier Displaying Component (Sub-ODC) [4].

Now replace the ϵ^{th} row of $\mathbf{1}$ with zero and then augment the $n-1\times p$ data set $\mathbf{Y}_{(\epsilon)}$ to $n\times p$ by replacing the ϵ^{th} row by a vector of zeros. The modified matrices $\mathbf{1}$ and $\mathbf{Y}_{(\epsilon)}$ are represented by $\mathbf{1}_{aug}$ and $\mathbf{Y}_{(\epsilon)aug}$, respectively. A projection of the mean-corrected data $\mathbf{Y}_{(\epsilon)}$ on the Sub-ODC is taken to obtain a univariate data

$$\mathbf{t}_2 = (\mathbf{Y}_{(\epsilon)aug} - (\mathbf{1}_{aug}\times\bar{\mathbf{y}}'))\times\varphi \tag{5}$$

The data set $\mathbf{T}_{n\times 2} = [\mathbf{t}_1\ \mathbf{t}_2]$ gives a bivariate equivalent of $\mathbf{X}$. Thus, $\mathbf{T}_{n\times 2}$ provides a dimensionally reduced data set derived from the original data $\mathbf{X}$. A scatter plot of $\mathbf{T}_{n\times 2}$ may be generated in the ODC-Sub-ODC plane.

Illustration In Figure 2.1 we have the scatter plot of the bivariate equivalent of the Artificial data set described in the previous section. It can be seen that observation 34 is extreme along the 1-ODC dimension (and is thus the single outlier). When observation 34 is deleted, the observation with the greatest generalized distance in the reduced data ($\mathbf{Y}_{(34)}$) is observation 45 which is extreme on the Sub-ODC. Figure 2.1 shows that observations 9 and 22 could also be candidates for an outlying pair.

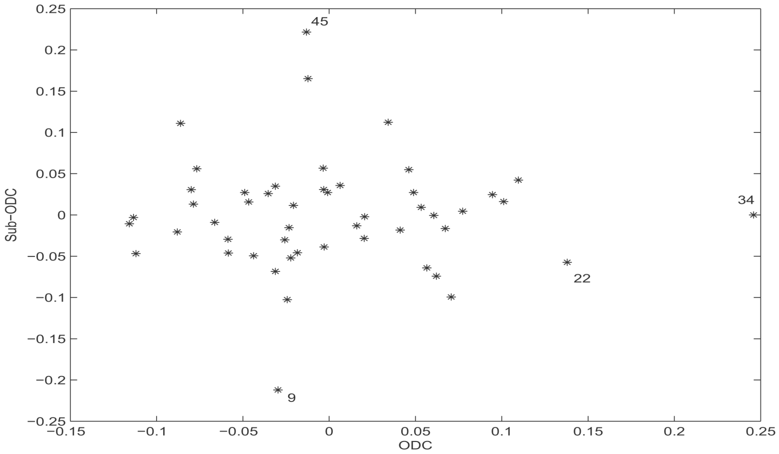

Figure 2.1
Scatter-plot of Artificial Data in ODC-Sub-ODC Plane

After describing the next two phases of the method, it will then be possible to identify the two most extreme pair of observations in this dataset.

2.2 The Least Squares Ellipse and its Orientation

This phase involves two main steps: (1) The determination of a least squares ellipse that fits the scatter of points, and (2) The determination of the orientation of the ellipse and its centre.

2.2.1 Determination of Least Squares Ellipse

An ellipse is a special case of a general conic which can be described by an implicit second order polynomial

$$E(x,y) = \begin{pmatrix} x & y & 1 \end{pmatrix} \begin{pmatrix} a & \frac{b}{2} & \frac{d}{2} \\ \frac{b}{2} & c & \frac{e}{2} \\ \frac{d}{2} & \frac{e}{2} & f \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = 0 \tag{6}$$

with an ellipse-specific constraint

$$b^2 - 4ac < 0. \tag{7}$$

The coefficients of the ellipse a, b, c, d, e, f are the parameters of ellipse and (x, y) are coordinates of points on it. The polynomial $E(x, y)$ is the algebraic distance of the point (x, y) to the given conic. To fit a general ellipse to a set of data points $(x_i, y_i),\ i = 1, 2, \cdots, n$, is to minimize the sum of squared algebraic distances of (x_i, y_i) to the conic subject to the constraint in Equation (7). Under a proper scaling, the inequality constraint, Equation (7), can be changed into an equality constraint $4ac - b^2 = 1$. Let $\mathbf{a}' = (a, b, c, d, e, f)$ and$\mathbf{x}' = (x^2, xy, y^2, x, y, 1)$. The ellipse-specific fitting problem may be formulated as

$$\min_{\mathbf{a}} ||\mathbf{Da}||^2 \qquad \text{subject to} \qquad \mathbf{a}'\mathbf{Ca} = 1 \tag{8}$$

where $\mathbf{C}_{n\times 6}$ is the constraint matrix and $\mathbf{D}_{n\times 6}$ is the design matrix.

Now let $\mathbf{S}_e = \mathbf{D}'\mathbf{D}$. By introducing the Lagrange Multiplier λ, we write the formulations in Equation (8) as $\psi = \mathbf{a}'\mathbf{S}_e\mathbf{a} + \lambda(1 - \mathbf{a}'\mathbf{Ca})$. Hence, differentiating ψ with respect to $\mathbf{a}$ and equating to zero, the conditions for optimal solution of $\mathbf{a}$ are obtained as

$$\begin{aligned} \mathbf{S}_e\mathbf{a} &= \lambda\mathbf{Ca} \\ \mathbf{a}'\mathbf{Ca} &= 1 \end{aligned} \tag{9}$$

To determine the parameters, partitioning of the matrices in Equation (9) have been suggested in various ways (e.g. Halir & Flusser, 2000; Harker, O'leary & Zsombor-Murray, 2004). Going by the partitioning of [5], we write the design matrix, $\mathbf{D}$, into two $n \times 3$ matrices as follows:

$$\mathbf{D}_1 = \begin{pmatrix} x_1^2 & x_1y_1 & y_1^2 \\ x_2^2 & x_2y_2 & y_2^2 \\ \vdots & \vdots & \vdots \\ x_i^2 & x_iy_i & y_i^2 \\ \vdots & \vdots & \vdots \\ x_n^2 & x_ny_n & y_n^2 \end{pmatrix}, \qquad \mathbf{D}_2 = \begin{pmatrix} x_1 & y_1 & 1 \\ x_2 & y_2 & 1 \\ \vdots & \vdots & \vdots \\ x_i & y_i & 1 \\ \vdots & \vdots & \vdots \\ x_n & y_n & 1 \end{pmatrix} \tag{10}$$

The scatter matrix $\mathbf{D}'\mathbf{D}$ is then given as a partitioned matrix as

$$\mathbf{D}'\mathbf{D} = \left(\begin{array}{c|c} \mathbf{D}_1'\mathbf{D}_1 & \mathbf{D}_1'\mathbf{D}_2 \\ \hline \mathbf{D}_1'\mathbf{D}_2 & \mathbf{D}_2'\mathbf{D}_2 \end{array}\right)$$

For easy representation of the element matrices of $\mathbf{D}'\mathbf{D}$, which is represented by$\mathbf{S}_e$, let $\mathbf{D}_1'\mathbf{D}_1 = \mathbf{S}_{11}$, $\mathbf{D}_1'\mathbf{D}_2 = \mathbf{S}_{12}$, and $\mathbf{D}_2'\mathbf{D}_2 = \mathbf{S}_{22}$. Similarly, the constraint matrix $\mathbf{C}$ (Equation (9)) can be expressed as

$$\mathbf{C} = \left(\begin{array}{c|c} \mathbf{C}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right), \qquad \text{where} \qquad \mathbf{C}_1 = \begin{pmatrix} 0 & 0 & 2 \\ 0 & -1 & 0 \\ 2 & 0 & 0 \end{pmatrix} \tag{11}$$

and $\mathbf{0}$ is a 3×3 zero matrix. The vector of coefficients $\mathbf{a}$ is split accordingly as

$$\mathbf{a} = \left(\begin{array}{c} \mathbf{a}_1 \\ \hline \mathbf{a}_2 \end{array}\right), \quad \text{where} \quad \mathbf{a}_1 = \begin{pmatrix} a \\ b \\ c \end{pmatrix} \text{ and } \mathbf{a}_2 = \begin{pmatrix} d \\ e \\ f \end{pmatrix} \tag{12}$$

Using these partitions, the first of the conditions in Equation (9) becomes

$$\left(\begin{array}{c|c} \mathbf{S}_{11} & \mathbf{S}_{12} \\ \hline \mathbf{S}_{21}' & \mathbf{S}_{22} \end{array}\right)\left(\begin{array}{c} \mathbf{a}_1 \\ \hline \mathbf{a}_2 \end{array}\right) = \lambda\left(\begin{array}{c|c} \mathbf{C}_1 & \mathbf{0} \\ \hline \mathbf{0} & \mathbf{0} \end{array}\right)\left(\begin{array}{c} \mathbf{a}_1 \\ \hline \mathbf{a}_2 \end{array}\right)$$

Hence, we obtain the equations

$$\begin{aligned} \mathbf{S}_{11}\mathbf{a}_1 + \mathbf{S}_{12}\mathbf{a}_2 &= \lambda\mathbf{C}_1\mathbf{a}_1 \\ \mathbf{S}_{21}^{'}\mathbf{a}_1 + \mathbf{S}_{22}\mathbf{a}_2 &= \mathbf{0} \end{aligned} \tag{13}$$

Before proceeding to solve the system of equations, there is a condition to examine for the existence of real solution of the system. The matrix

$$\mathbf{S}_{22} = \begin{pmatrix} \Sigma x_i^2 & \Sigma x_i y_i & \Sigma x_i \\ \Sigma x_i y_i & \Sigma y_i^2 & \Sigma y_i \\ \Sigma x_i & \Sigma y_i & n \end{pmatrix}$$

is exactly a scatter matrix of fitting a line through a set of data points. It is known (Haralick & Shapiro, 1993) that this matrix is singular only if all the points lie on a line. In this study, the points (x_i, y_i) are generated by projections of p-dimensional dataset onto the ODC and the Sub-ODC. By the properties of the two dimensions already described, the scatter of points in the plane provided by these two dimensions cannot lie on a line. Thus, $\mathbf{S}_{22}$ is regular in this case. Consequently, there is a solution for fitting an ellipse through points in the ODC-Sub-ODC plane.

Returning to the solution of $\mathbf{a}^{'} = (\mathbf{a}_1, \mathbf{a}_2)$, from the second of Equation (13),

$$\mathbf{a}_2 = -\mathbf{S}_{22}^{-1}\mathbf{S}_{21}^{'}\mathbf{a}_1 \tag{14}$$

Substituting this into the first of Equation (13) gives

$$(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}^{'})\mathbf{a}_1 = \lambda\mathbf{C}_1\mathbf{a}_1$$

and since $\mathbf{C}_1$ is regular,

$$\begin{aligned} \mathbf{C}_1^{-1}(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}^{'})\mathbf{a}_1 &= \lambda\mathbf{a}_1 \\ \left(\mathbf{C}_1^{-1}(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}^{'}) - \mathbf{I}\lambda\right)\mathbf{a}_1 &= 0 \end{aligned}$$

It follows therefore that $\mathbf{a}_1$ is the eigenvector of the reduced scatter matrix

$$\mathbf{M} = \mathbf{C}_1^{-1}(\mathbf{S}_{11} - \mathbf{S}_{12}\mathbf{S}_{22}^{-1}\mathbf{S}_{21}^{'}) \tag{15}$$

of size 3 × 3 and λ is the corresponding eigenvalue. The conditions of Equation (9) can now be restated as

$$\begin{aligned} \mathbf{M}\mathbf{a}_1 &= \lambda\mathbf{a}_1 \\ \mathbf{a}_1^{'}\mathbf{C}_1\mathbf{a}_1 &= 1 \\ \mathbf{a}_2 &= -\mathbf{S}_{22}^{-1}\mathbf{S}_{21}^{'}\mathbf{a}_1 \\ \mathbf{a} &= \begin{pmatrix} \mathbf{a}_1 \\ \mathbf{a}_2 \end{pmatrix} \end{aligned} \tag{16}$$

MATLAB codes have been outlined by [5] for computing the ellipse coefficients.

2.2.2 Derivation of the Orientation and Centre of the Ellipse

The coefficient vector $\mathbf{a}$ in Equation (16) gives a tilted ellipse. Suppose that $\mathbf{a}^{'} = (a_{11}, a_{12}, a_{13}, a_{21}, a_{22}, a_{23})$, and let a transformation of points (x_i, y_i) in the ODC-Sub-ODC plane be given by $T : x^{'} = cx + sy; \quad y^{'} = -sx + cy$, where $c = \cos\phi$ and $s = \sin\phi$. If this transformation should return a non-tilted ellipse, then the angle ϕ must be the required angle of orientation of the ellipse. Substituting this transformation in the equation $E = \mathbf{x}\cdot\mathbf{a}$, we obtain the expression $E_n = \mathbf{x}\cdot\mathbf{a}^*$, where the components of $\mathbf{a}^* = (a_{11}^*, a_{12}^*, a_{13}^*, a_{21}^*, a_{22}^*, a_{23}^*)$

are such that, for example, $a^*_{12} = 2a_{11}cs + (c^2 - s^2)a_{12} - 2a_{13}cs$. For a non-tilt ellipse, $a^*_{12} = 0$. After a few simplifications the orientation of the ellipse in Equation (16) is given as

$$\phi = \frac{1}{2}\tan^{-1}\Big(\frac{a_{12}}{a_{13} - a_{11}}\Big) \tag{17}$$

The coefficients of the non-tilt ellipse are then obtained by substituting sinϕ and cosϕ into the components of $\mathbf{a}^*$. By this substitution, the non-tilt ellipse is given by the coefficients $\mathbf{a}^* = (a^*_{11}, a^*_{13}, a^*_{21}, a^*_{22}, a^*_{23})'$, where $a^*_{23} = a_{23}$.

It may be necessary to identify the centre of the main cloud of the data points. To determine this centre, (i.e. of the non-tilt ellipse) we express the equation $a^*_{11}x^2 + a^*_{13}y^2 + a^*_{21}x + a^*_{22}y + a^*_{23} = 0$ by square completion as

$$\frac{\Big(x + \dfrac{a^*_{21}}{2a^*_{11}}\Big)^2}{\dfrac{a^{**}_{23}}{a^*_{11}}} + \frac{\Big(y + \dfrac{a^*_{22}}{2a^*_{13}}\Big)^2}{\dfrac{a^{**}_{23}}{a^*_{13}}} = 1$$

where $a^{**}_{23} = -a_{23} + \dfrac{a^{*2}_{21}}{4a^*_{11}} + \dfrac{a^{*2}_{22}}{4a^*_{13}}$. In this form, the centre (C) of the non-tilt ellipse is given as $\Big(-\dfrac{a^*_{21}}{2a^*_{11}}, -\dfrac{a^*_{22}}{2a^*_{13}}\Big)$. Lastly, the centre of the tilt ellipse is obtained by $T^{-1}C$.

2.3 Determination of a Projection Vector

We first note that in the plane provided by the ODC and the Sub-ODC, the most extreme observation along the ODC dimension, $\mathbf{x}_\epsilon$, has coordinates $\mathbf{x}_\epsilon\big(U(\mathbf{x}_\epsilon, \mathbf{S}),\ 0\big)$. Given the orientation ϕ of the minimum ellipse, a unit vector that is perpendicular to this orientation is $\mathbf{v} = (\sin\phi,\ -\cos\phi)$. We choose a point $\mathbf{x}^B_\epsilon\big(U(\mathbf{x}_\epsilon, \mathbf{S}),\ k\big)$, directly below $\mathbf{x}_\epsilon$, where $k \leq 0$ is chosen such that k is equal to the smallest value along the Sub-ODC dimension. Now a vector that passes through $\mathbf{x}^B_\epsilon$ and parallel to $\mathbf{v}$ is given by

$$\mathbf{w}_e = \begin{pmatrix} k\tan\phi + U(\mathbf{x}_\epsilon, \mathbf{S}) \\ -k - \dfrac{U(\mathbf{x}_\epsilon, \mathbf{S})}{\tan\phi} \end{pmatrix} \tag{18}$$

This vector, which is perpendicular to the orientation of the ellipse, will subsequently be referred to as the *Perpendicular Elliptical Vector*. This is to distinguish it from the *Parallel Elliptical Vector* given by

$$\mathbf{u}_e = \begin{pmatrix} \dfrac{k}{\tan\phi} + U(\mathbf{x}_\epsilon, \mathbf{S}) \\ -k - U(\mathbf{x}_\epsilon, \mathbf{S})\tan\phi \end{pmatrix} \tag{19}$$

which passes through $\mathbf{x}^B_\epsilon$ and parallel to the orientation of the ellipse.

2.4 Some Rules for Outlier Detection

The use of the *Elliptical* vector in Equation (18) (or Equation (19)) involves a number of heuristics to successfully identify the pair of outliers. The following are three properties of the observations that constitute the pair of outliers after projection on the elliptical vector:

1. An observation that sticks out distinctly at one end of the (perpendicular) projection must necessarily be one of the outlying pair.

2. Two observations that stick out distinctly on either ends of the (perpendicular) projection constitute the pair of outliers.

3. If only one observation sticks out distinctly at one end of the (perpendicular) projection, but the other member of the pair does not stick out clearly on the other end, then the reduced two-dimensional data may be projected on the parallel elliptical vector. Each of the observations that stick out on the two separate projections then combine to form the pair of outliers.

3. ILLUSTRATION OF THE METHOD

We illustrate the technique by completing the detection of a pair of outliers in the Artificial data in Section 2. Figure 3.1 is the one-dimensional plot of the projection of the reduced data displayed in Figure 2.1 on the perpendicular elliptical vector.

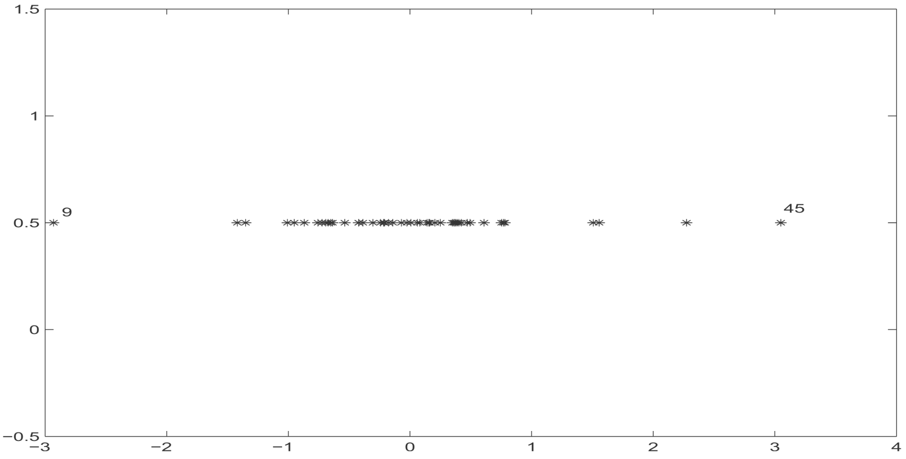

Figure 3.1
Projection of Artificial Data on Elliptical Vector

From Figure 3.1, it is observed that observations 9 and 45 are extreme on either ends of the projection. According to the second of our rules outlined above, 9 and 45 must be the pair of outliers. In fact, when we test for two outliers by calculating Wilk's two-outlier statistic, $r_2 = \min_T R_T$, the pair (9, 45) has the smallest value of r_2 (= 0.5656). The identification of these two observations as the pair of outliers shows that the most outlying pair in a dataset does not always include the single outlier.

4. A BACK-UP PROCEDURE FOR HIGHER DIMENSIONAL DATASET

To successfully detect the right pair of outliers in data of dimension greater than 3, we further propose a 'back-up' procedure that could be implemented after the pair of outliers have been detected in a first round implementation of the proposed method. Now the incidence of an outlying observation is as a result of a marked deviation in the observation along one or more dimensions on which data were generated. Thus, a dimension along which there is smallest variation would not contribute significantly to outlier detection. Therefore, to assess the correctness of a detected pair $(\mathbf{x}_\epsilon, \mathbf{x}_\iota)$ as outliers, we delete the dimension along which there is smallest variation, and then implement the procedure again. The pair of outliers that emerge in this second round of implementation are the right pair of outliers.

Table 1 shows single outliers and a pair of outliers detected by means of the generalized distance (Gen. Dist.) from the sample mean, the proposed method and the Wilk's ratio method in some widely studied data sets of varying sizes, n and dimensionality, p. These data sets include the Milk Transportation Cost Data (Johnson & Wichern, 2002), the three types of Iris datasets (Johnson & Wichern, 2002; Anderson, 2003) and the US Food Price data (Sharma, 1996).

Table 1
Performances of Proposed Method and Two Others

	Single	Pair of outliers detected by		
Data Set	Outlier	Gen. Dist.	Proposed	Wilk's Ratio
Artificial data ($p = 4, n = 50$)	34	34, 45	9, 45	9, 45
Iris Setosa ($p = 4, n = 50$)	42	42, 44	42, 23	42, 23
Iris Versicolor ($p = 4, n = 50$)	19	19, 49	19, 49	19, 49
Iris Virginica ($p = 4, n = 50$)	19	19, 32	19, 18	19, 32
Transport-Cost ($p = 3, n = 36$)	9	9, 21	9, 21	9, 21
U.S Food Price ($p = 5, n = 23$)	10	10, 1	10, 16	10, 16

Table 1 shows that with exception of the Iris Virginica, the results of both the proposed method and the Wilk's ratio method coincide in all the samples used. Also, with exception of the Artificial dataset, all others (which are real datasets) include the single outlier in the pair of outliers.

5. CONCLUSION

The paper has considered a procedure for detecting a pair of outliers in multivariate data. We discover that if two observations constitute a pair of outliers, then there must exist a one-dimensional projection (given by the *Perpendicular Elliptical Vector*) on which the pair is extreme at either ends and separated clearly from the remaining observations. If the two outliers are not distinct on such a one-dimensional projection, a combination of two separate one-dimensional projections (the other given by the *Parallel Elliptical Vector*) may be obtained. Each of the observations that is extreme on the two separate projections combine to form the pair of outliers.

The method proposed aims at reducing the dimensionality of the original data and then removes the influence of the generalised sample mean by eliminating all observations that are close to it.

REFERENCES

[1] Anderson, T. W. (2003). *Introduction to Multivariate Statistical Analysis*. New Jersey, Prentice Hall.

[2] Caroni, C. & Prescott, P. (1992). Sequential Application of Wilk's Multivariate Outlier Test. *Applied Statistics*, *41*, 355-364.

[3] Gordor, B. K. & Fieller, N. R. J. (1999). How to Display an Outlier in Multivariate Datasets. *Journal of Applied Sciences & Technology*, *4*(2).

[4] Halir R. & Flusser, J. (2000). *Numerically Stable Direct Least Squares Fitting of Ellipses*. Technical Report citeseer.nj.nec.com/350661.html, Department of Software Engineering, Charles University, Czech Republic.

[5] Haralick R. M. & Shapiro, L., G. (1993). *Computer and Robot Vision*, *1*. Addison Wesley.

[6] Harker, M., O'leary P. & Zsombor-Murray, P. (2004). *Direct and Specific Fitting of Conics to Scattered Data*. Technical Report, Institute for Automation, University of Leoben.

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

[8] Rousseeuw, P. J. (1985). Multivariate Estimation with High Breakdown Point. *Mathematical Statistics and Application*, *B*, 283-297.

[9] Rousseeuw, P. J. & Leroy, A. (1987). *Robust Regression and Outlier Detection.* New York: Wiley.

[10] Sharma, S. (1996). *Applied Multivariate Techniques*. New York: Wiley.

Appendix: Artificial Data

No.	X_1	X_2	X_3	X_4
1	0.074	0.037	-0.479	0.296
2	-1.051	-0.095	0.098	-1.495
3	1.767	-1.469	-0.592	2.392
4	-2.623	0.726	-1.317	0.106
5	0.281	-0.650	-2.044	0.827
6	1.717	-0.429	0.457	0.825
7	-1.012	0.504	-0.514	0.244
8	1.404	-0.549	0.558	1.473
9	1.804	-1.849	-1.477	-0.921
10	-1.423	-0.566	-0.342	0.120
11	0.560	-0.188	1.035	0.242
12	0.605	1.324	-0.460	0.194
13	0.480	1.954	-0.254	-0.213
14	-1.166	1.151	-1.395	-0.438
15	1.203	1.511	1.219	-1.623
16	-1.135	-0.541	-0.392	1.028
17	-0.027	0.742	-0.784	0.151
18	0.580	-0.575	-0.234	0.118
19	-1.111	0.952	-0.596	-0.733
20	-0.873	-1.631	-0.423	1.627
21	1.315	-0.775	0.908	0.025
22	0.957	0.191	-1.607	1.344
23	-0.736	-0.494	-0.177	-0.738
24	-2.527	-0.661	-2.139	-0.391
25	-0.352	-1.381	-1.146	0.801
26	-0.775	0.970	-1.420	0.169
27	0.589	1.026	0.331	-1.977
28	-0.153	0.064	0.354	-0.306
29	1.808	-1.478	0.886	1.677
30	1.720	1.263	1.020	-0.417
31	-1.050	0.147	1.306	2.716
32	-1.040	0.152	1.341	0.529
33	0.058	1.614	0.208	-0.524
34	1.179	1.914	-1.744	1.294
35	-0.096	0.353	1.241	-0.584
36	-0.564	-0.846	-0.892	0.070
37	0.356	-0.952	-1.107	-0.354
38	-1.570	0.409	-0.141	-1.143
39	0.470	-1.022	0.571	-0.772
40	-0.480	0.972	-1.112	0.798
41	0.222	-0.687	0.672	1.464
42	0.397	-0.003	0.444	0.160
43	-0.117	0.175	-0.758	-0.681
44	-1.359	0.535	-2.054	0.335
45	-1.640	1.721	1.726	1.181
46	0.116	-0.400	0.975	3.444
47	-1.391	-1.200	-0.305	0.088
48	0.726	-0.079	-1.076	0.284
49	1.149	0.126	0.193	-1.171
50	-0.640	1.464	-1.224	-1.618