Journal of Informatics and Mathematical Sciences
Volume 4 (2012), Number 2, pp. 229–239
© RGN Publications
http://www.rgnpublications.com

On the One-Outlier Displaying Component

B.K. Nkansah and B.K. Gordor

Abstract. A method of displaying an outlier in a multivariate data set is the Outlier Displaying Component. This method is based on the sample mean vector and the sum of squares and cross- product matrix. The main weakness of this method is that both of these measures involve the very outlier that is being detected. This paper presents an approach to eliminating this weakness. By eliminating the outlier from the sample mean vector and the sum of squares and cross-product matrix, the proposed method combines a number of advantages: It enhances the separation of the outlier from the rest of the data so that it appears more distinct. It also increases the general dispersion in the projected data so that the presence of multiple outliers could be revealed.

1. Introduction

A way of obtaining a revealing view of a multivariate dataset is to find its univariate equivalent by a projection vector. One of the projection methods that specifically seeks to highlight the outlier so that it 'sticks out' from the remaining observations is the One-Outlier Displaying Component [7]. It is well known that the observation that has the most distinctly projected univariate value is always the one with the largest Mahalanobis distance from the general sample mean. Equivalently, the single outlier, $\mathbf{x}_\epsilon$, among a p-dimensional data set, $\mathbf{x}_{n\times p} = (x_1, x_2, \cdots, x_n)'$, is the one for which the Wilk's ratio

$$r_1 = \frac{|\mathbf{S}_{(\epsilon)}|}{|\mathbf{S}|} \tag{1}$$

is minimum, where $\mathbf{S}$ is the sample sum of squares and cross product (SSCP) matrix and $\mathbf{S}_{(\epsilon)}$, is the SSCP matrix of the remaining $(n-1)$ observations when the outlier is deleted from the sample. The matrices, $\mathbf{S}$ and $\mathbf{S}_{(\epsilon)}$, are given respectively, by

$$\mathbf{S} = \sum_{j=1}^{n}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})', \quad \text{where } \bar{\mathbf{x}} = \frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_j, \tag{2}$$

Key words and phrases. Outliers; Outlier displaying component; Outlier detection.

and

$$\mathbf{S}_{(\epsilon)} = \sum_{j \notin I} (\mathbf{x}_j - \bar{\mathbf{x}}_{(I)})(\mathbf{x}_j - \bar{\mathbf{x}}_{(I)})', \quad \text{where } \bar{\mathbf{x}}_{(I)} = \frac{1}{n-k} \sum_{j \notin I} \mathbf{x}_j \ . \tag{3}$$

The vector $\bar{\mathbf{x}}$ is the general sample mean and the vector $\bar{\mathbf{x}}_{(I)}$ is the mean of the remaining $(n-1)$ observations when the outlier is deleted from the data. The set I is an indexed set of outliers in the sample which, in the case of a single outlier, contains only one element, ϵ. The ratio in Equation (1) has been expressed ([7]; [4]) in various ways to be equivalent to

$$r_1 = 1 - \frac{n}{n-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \ . \tag{4}$$

Thus, the outlier, $\mathbf{x}_\epsilon$, is that observation for which r_1 is minimum. Equivalently, the outlier is that for which

$$U(\bar{\mathbf{x}}, \mathbf{S}) = (\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \ . \tag{5}$$

is maximum.

It is evident that for the purpose of detecting a single outlier, the general sample mean, $\bar{\mathbf{x}}$, and hence, the sample SSCP matrix, $\mathbf{S}$, are appropriate measures of average and dispersion. However, the problems associated with these measures are well known. One problem is that, $\bar{\mathbf{x}}$ and $\mathbf{S}$ are themselves influenced by the outlier.

In this paper, we analyse the projection approach of [7] to the detection and display of a single outlier. Subsequently, we characterize the drawback of the approach and propose an improved method. We first describe the original approach. Then based on what we will refer to as the *difference decomposition*, an alternative projection method is proposed. We will then show analytically that the proposed method is able to project the outlier observation more distinctly than the original method.

2. The One-Outlier Displaying Component

In this section, we provide a review of the One-Outlier Displaying Component (1-ODC). Following work by [6] on outliers in q-dimensional projections of the p-dimensional samples $(q < p)$, [7] derived a projection vector, β, which converts a p-dimensional observation, $\mathbf{X}_{n \times p}$, into a corresponding univariate observations, y_i; $1, 2, \cdots, n$ such that

$$y_i = \beta' \mathbf{x}_i \ .$$

An equivalent expression for $U(\bar{\mathbf{x}}, \mathbf{S})$ in Equation (5), of the distance of y_ϵ after projection of $\mathbf{x}_\epsilon$ then becomes

$$U(\bar{\mathbf{x}}, \mathbf{S}; \beta) = \frac{\beta'(\mathbf{x}_\epsilon - \bar{\mathbf{x}})(\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\beta}{\beta' \mathbf{S} \beta} \ . \tag{6}$$

The vector, β, is chosen to maximize Equation (6) subject to the constraint that $\beta'\mathbf{S}\beta = c$, where c is an arbitrary constant. The solution of the maximization problem reveals that β is the eigenvector associated with the $p \times p$ matrix

$$\mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}})(\mathbf{x}_\epsilon - \bar{\mathbf{x}})' .$$

The solution further shows that β is given by

$$\beta = \mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \tag{7}$$

and the associated eigenvalue was found to be the squared Mahalanobis distance given in Equation (5). This vector provides the dimension on which the labelled outlier, $\mathbf{x}_\epsilon$, sticks out the most from the remaining observations and is referred to as the 1-ODC.

If we substitute the vector $\mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}})$ for β in Equation (6), it can be shown that

$$U(\bar{\mathbf{x}}, \mathbf{S}; \beta) = U(\bar{\mathbf{x}}, \mathbf{S}) . \tag{8}$$

Since $U(\bar{\mathbf{x}}, \mathbf{S})$ is the likelihood based statistic ([2]; [4]) for testing the extreme p-dimensional observation, $\mathbf{x}_\epsilon$, in the original sample, this result shows that the value of the discordancy test statistic in (the original) p dimensions is numerically the same as that in the single dimension provided by the 1-ODC.

3. **A Modification of the One-Outlier Displaying Component**

Even though the original 1-ODC approach is theoretically correct, its ability to isolate the outlier can still be enhanced. The main drawback in its performance is due to the use of the sample mean in the projection vector. As has been pointed out earlier, the computation of the sample mean is itself influenced by the outlier. As a result, its involvement in the detection and display of the outlier can negatively affect this effort. In this section, we obtain a modification of the projection vector that excludes the use of the sample mean.

First, we examine the difference, $\mathbf{x}_j - \bar{\mathbf{x}}$, between any observation $\mathbf{x}_j$ and the general mean $\bar{\mathbf{x}}$. Let I_k denote the set of k outliers in the data set. Then, the general sample mean vector, $\bar{\mathbf{x}}$, is given by

$$\bar{\mathbf{x}} = \frac{(n-k)\bar{\mathbf{x}}_{(I_k)} + k\bar{\mathbf{x}}_{I_k}}{n},$$

implying that

$$n\bar{\mathbf{x}} = (n-k)\bar{\mathbf{x}}_{(I_k)} + k\bar{\mathbf{x}}_{I_k} . \tag{9}$$

Now, writing the left hand side of Equation (9) as $n\bar{\mathbf{x}} + n\mathbf{x}_j - n\mathbf{x}_j$, we have

$$\begin{aligned} n\bar{\mathbf{x}} + n\mathbf{x}_j - n\mathbf{x}_j &= (n-k)\bar{\mathbf{x}}_{(I_k)} + k\bar{\mathbf{x}}_{I_k} \\ n(\bar{\mathbf{x}} - \mathbf{x}_j) &= (n-k)\bar{\mathbf{x}}_{(I_k)} + k\bar{\mathbf{x}}_{I_k} - n\mathbf{x}_j . \end{aligned}$$

Multiplying through by -1, we have

$$\begin{aligned} n(\mathbf{x}_j - \bar{\mathbf{x}}) &= n\mathbf{x}_j - (n-k)\bar{\mathbf{x}}_{(I_k)} - k\bar{\mathbf{x}}_{I_k} \\ &= (n-k)\mathbf{x}_j + k\mathbf{x}_j - (n-k)\bar{\mathbf{x}}_{(I_k)} - k\bar{\mathbf{x}}_{I_k} \\ &= (n-k)\mathbf{x}_j - (n-k)\bar{\mathbf{x}}_{(I_k)} + k\mathbf{x}_j - k\bar{\mathbf{x}}_{I_k} \ . \end{aligned}$$

Therefore,

$$\mathbf{x}_j - \bar{\mathbf{x}} = \frac{n-k}{n}(\mathbf{x}_j - \bar{\mathbf{x}}_{(I_k)}) + \frac{k}{n}(\mathbf{x}_j - \bar{\mathbf{x}}_{I_k}) \ . \tag{10}$$

Equation (10) provides a partitioning of the difference $\mathbf{x}_j - \mathbf{x}$ into a weighted sum of two components in the presence of k outliers. These components are: (1) the difference between $\mathbf{x}_j$ and the mean, $\bar{\mathbf{x}}_{(I_k)}$, of the remaining $(n-k)$ observations which excludes the set of k outliers; and (2) the difference between $\mathbf{x}_j$ and the mean, $\bar{\mathbf{x}}_{I_k}$, of the set of outliers. In outlier detection, $n \gg k$, implying that $\frac{k}{n}$ is very close to zero. Consequently, if n is large, regardless of the relative position of $\mathbf{x}_j$, $\mathbf{x}_j - \bar{\mathbf{x}}$ is approximately equal to the first component as the second vanishes.

Particularly, in the single outlier case, $\bar{\mathbf{x}}_{I_k} = \mathbf{x}_\epsilon$, which is the labelled outlier, and Equation (10) reduces to

$$\mathbf{x}_\epsilon - \bar{\mathbf{x}} = \frac{n-1}{n}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \ . \tag{11}$$

Following the approach outlined in the previous section, we find the projections of the $p-$dimensional observations $\mathbf{x}_j$ into corresponding univariate observations y_j such that for some β_e, $y_\epsilon = \beta_e' \mathbf{x}_\epsilon$. Now the distance of y_ϵ from the remaining $n-1$ observations is

$$\begin{aligned} U(y_\epsilon; \bar{\mathbf{x}}_(\epsilon), \mathbf{S}_(\epsilon)) &= (y_\epsilon - \bar{y})' \mathbf{S}_y^{-1} (y_\epsilon - \bar{y}) \\ &= \frac{\beta_e'(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})\beta_e}{\beta_e' \mathbf{S}_{(\epsilon)} \beta_e} \ . \end{aligned} \tag{12}$$

If we maximize this expression subject to the constraint $\beta_e' \mathbf{S}_{(\epsilon)} \beta_e = c$, we obtain

$$\beta_e = \mathbf{S}_{(\epsilon)}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \ . \tag{13}$$

Subsequently,

$$U(\bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)}) = (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \ . \tag{14}$$

Now the that the generalized distance of $y_\epsilon = \beta' \mathbf{x}_\epsilon$, is given by

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

We now show that

$$(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \gg (\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \ .$$

Writing the inequality as an equation, we have

$$(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) = (\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}) + \kappa$$

where κ is a constant.

Thus,

$$(\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) = (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})'\mathbf{S}_{(\epsilon)}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) - \kappa \ . \tag{15}$$

We know that for any set I of k outliers, $\mathbf{S}$ and $\mathbf{S}_{(I)}$ are related by the equation

$$\mathbf{S} = \mathbf{S}_{(I)} + \mathbf{A}_{I_k} \ .$$

Taking the inverse of both sides, we obtain

$$\mathbf{S}^{-1} = (\mathbf{S}_{(I)} + \mathbf{A}_{I_k})^{-1} \ . \tag{16}$$

Thus, the left hand side of Equation (15) is written as

$$(\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) = (\mathbf{x}_\epsilon - \bar{\mathbf{x}})'(\mathbf{S}_{(I)} + \mathbf{A}_{I_k})^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \ . \tag{17}$$

Again, it has been shown [4] that in the case of a single outlier,

$$\mathbf{A}_{I_1} = \frac{n}{n-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}})(\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \tag{18}$$

which is of rank 1 and therefore has no inverse; $\mathbf{S}_{(I)}$ is of full rank and hence has an inverse. To obtain the inverse in Equation (16), we recall the Sherman-Morrison result ([10], [8], [5]) that if $\mathbf{G}$ and $\mathbf{G}+\mathbf{E}$ are non-singular matrices where $\mathbf{E}$ is a matrix of rank 1, then the inverse of the sum $\mathbf{G}+\mathbf{E}$ is

$$(\mathbf{G}+\mathbf{E})^{-1} = \mathbf{G}^{-1} - \frac{1}{1+tr\,\mathbf{E}\mathbf{G}^{-1}}\mathbf{G}^{-1}\mathbf{E}\mathbf{G}^{-1}$$

where $tr\,\mathbf{E}\mathbf{G}^{-1} \neq -1$. If we relate $\mathbf{S}_{(I_k)}$ to $\mathbf{G}$ and $\mathbf{A}_{(I_1)}$ to $\mathbf{E}$, we obtain

$$(\mathbf{S}_{(I)} + \mathbf{A}_{(I_1)})^{-1} = \mathbf{S}_{(I)}^{-1} - \frac{1}{1+tr\,\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}}\mathbf{S}_{(I)}^{-1}\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1} \ .$$

Substituting the result above into Equation (17) and noting the result in Equation (11), we obtain

$$\begin{aligned}
U(\bar{\mathbf{x}}, \mathbf{S}) &= (\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\left\{\mathbf{S}_{(I)}^{-1} - \frac{1}{1+tr\,\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}}\mathbf{S}_{(I)}^{-1}\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}\right\}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \\
&= (\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\mathbf{S}_{(I)}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) - \frac{1}{1+tr\,\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}}(\mathbf{x}_\epsilon - \bar{\mathbf{x}})'\{\mathbf{S}_{(I)}^{-1}\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}\}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \\
&= \left(\frac{n-1}{n}\right)^2 [(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})'\mathbf{S}_{(I)}^{-1}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \\
&\quad - \frac{1}{1+tr\,\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})'\{\mathbf{S}_{(I)}^{-1}\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}\}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})] \ .
\end{aligned} \tag{19}$$

If n is large, Equation (19) becomes approximately

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

where

$$\kappa = \frac{1}{1+tr\,\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})'\{\mathbf{S}_{(I)}^{-1}\mathbf{A}_{I_1}\mathbf{S}_{(I)}^{-1}\}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})$$

implying that

$$(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \gg (\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \mathbf{S}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}) \, .$$

Equation (19) also shows that the distance of the outlier from the mean $\bar{\mathbf{x}}$ is much less than its distance from the centre of the sample without the outlier. As a result, a projection of the data on the vector given in Equation (13) will reveal the outlier better. In the next section, we illustrate the performances of the proposed 1-ODC and the original 1-ODC using some data sets.

4. Illustration of the Performance of the Original and Modified 1-ODCs

Three sets of data have been used to illustrate the performance of the two projections discussed. These datasets are (1) the well studied Iris Setosa data [1] obtained from 50 plants on four dimensions: sepal length, sepal width, petal length and petal width; (2) the Milk Transportation-Cost data [9] obtained from 36 farms on three dimensions: fuel, repair and capital (all measured on a per-mile basis); (3) the U.S.A. Food-Price data [11] collected from 23 cities on five dimensions: bread, burger, milk, oranges and tomatoes (all measured in cents per pound). The third set of data has been attached in the appendix for reference since it is not as popular as the other two.

In Figures 1, 2 and 3, the projection unto the original 1-ODC is shown below the projection unto the modified 1-ODC. It can be seen from all the graphs that the projection unto the proposed 1-ODC (β_e) increases the spread in the projected data more than that achieved by the original 1-ODC (β). In particular, the distance of the specified outlier from the next isolated observation is much more greater with (β_e) than (β). In addition, as a result of the increase in dispersion, the proposed method reveals other observations that could be examined for outlyingness. For example, in Figure 1, observation 33 may be the next to consider as an outlier apart from observation 42. In Figure 2, observation 21 is more isolated from the observation on the left than observation 36 on the right. It is therefore not surprising that observations 9 and 21 have been identified [3] as the pair of outliers in this dataset.

5. Relative Efficiency of the Modified 1-ODC over the Original 1-ODCs

In Equation (19), we have shown that

$$\begin{aligned} U(\bar{\mathbf{x}}, \mathbf{S}) = \left(\frac{n-1}{n}\right)^2 & [(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(I)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \\ & - \frac{1}{1 + tr\, \mathbf{A}_{I_1} \mathbf{S}_{(I)}^{-1}} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \{\mathbf{S}_{(I)}^{-1} \mathbf{A}_{I_1} \mathbf{S}_{(I)}^{-1}\} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})] \, . \end{aligned} \tag{20}$$

We also know from Equation (18) that

$$\mathbf{A}_{I_1} = \frac{n}{n-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}})(\mathbf{x}_\epsilon - \bar{\mathbf{x}})' \, .$$

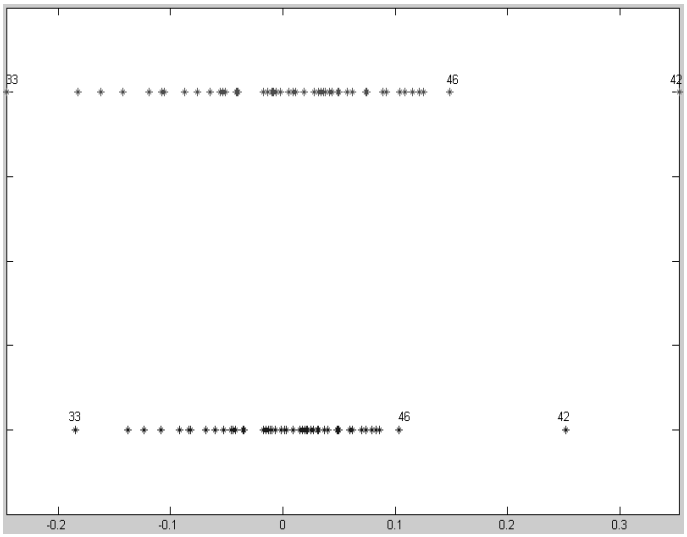

Figure 1. Projection of Iris Setosa on Original (below) and Modified (above) 1-ODCs

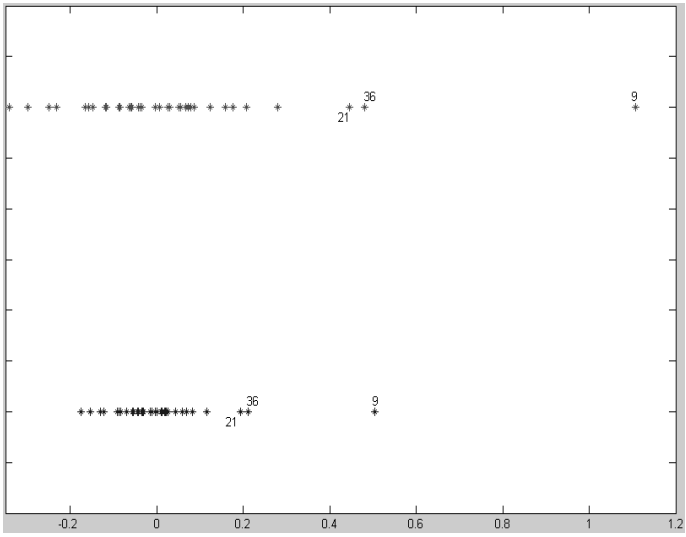

Figure 2. Projection of Transport data on Original (below) and Modified (above) 1-ODCs

Using a substitution from Equation (10), we obtain

$$\begin{aligned}\mathbf{A}_{I_1} &= \frac{n}{n-1}\left(\frac{n-1}{n}\right)^2(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \\ &= \frac{n-1}{n}(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \,.\end{aligned}$$

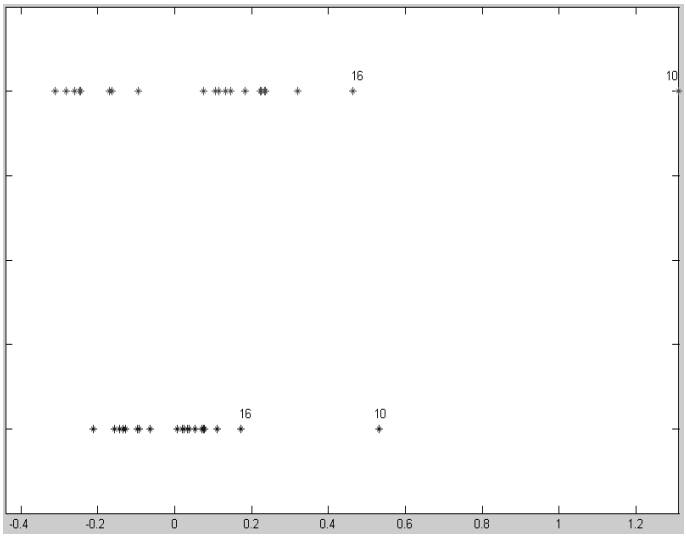

Figure 3. Projection of Food Price data on Original (below) and Modified (above) 1-ODCs

If we substitute this expression for $\mathbf{A}_{I_1}$ in Equation (20), we obtain the right hand side as

$$\begin{aligned}
&\left(\frac{n-1}{n}\right)^2 [(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \\
&- \lambda (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \{\mathbf{S}_{(\epsilon)}^{-1} [(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})'] \mathbf{S}_{(\epsilon)}^{-1}\} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})] \\
&\quad = \left(\frac{n-1}{n}\right)^2 [(\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)}) \\
&\qquad - \lambda (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})' \mathbf{S}_{(\epsilon)}^{-1} (\mathbf{x}_\epsilon - \bar{\mathbf{x}}_{(\epsilon)})] \\
&\quad = \left(\frac{n-1}{n}\right)^2 U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)}) [1 - \lambda U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)})]
\end{aligned}$$

where

$$\lambda = \frac{n-1}{n} \frac{1}{1 + tr\, \mathbf{A}_{I_1} \mathbf{S}_{(\epsilon)}^{-1}}, \qquad tr\, \mathbf{A}_{I_1} \mathbf{S}_{(I)}^{-1} \neq -1 .$$

Equation (20) now simplifies as

$$U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}, \mathbf{S}) = \left(\frac{n-1}{n}\right)^2 U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)}) [1 - \lambda U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)})] .$$

Therefore, we obtain the ratio

$$\frac{U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}, \mathbf{S})}{U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)})} = \left(\frac{n-1}{n}\right)^2 [1 - \lambda U(\mathbf{x}_\epsilon; \bar{\mathbf{x}}_{(\epsilon)}, \mathbf{S}_{(\epsilon)})]$$

or

$$\frac{U(\mathbf{x}_\epsilon;\bar{\mathbf{x}}_{(\epsilon)},\mathbf{S}_{(\epsilon)})}{U(\mathbf{x}_\epsilon;\bar{\mathbf{x}},\mathbf{S})}=\left(\frac{n}{n-1}\right)^2[1-\lambda U(\mathbf{x}_\epsilon;\bar{\mathbf{x}}_{(\epsilon)},\mathbf{S}_{(\epsilon)})]^{-1}\,. \tag{21}$$

Equation (21) gives the amount by which the distance of $\mathbf{x}_\epsilon$ from $\bar{\mathbf{x}}_{(\epsilon)}$ exceeds its distance from $\bar{\mathbf{x}}$. It thus, measures the relative efficiency of the modified method over the original. Now, since the ratio is maximum among all observations $\mathbf{x}_i$, it means the expression in the square brackets on the right of Equation (21) which we represent as

$$r_1^{(\epsilon)}=1-\lambda(\mathbf{x}_\epsilon-\bar{\mathbf{x}}_{(\epsilon)})'\mathbf{S}_{(\epsilon)}^{-1}(\mathbf{x}_\epsilon-\bar{\mathbf{x}}_{(\epsilon)}) \tag{22}$$

is smallest. In terms of $\bar{\mathbf{x}}_{(\epsilon)}$ and $\mathbf{S}_{(\epsilon)}$, Equation (22) may be seen to correspond to the one in Equation (4).

In the following table, we use Equation (21) to compute the relative efficiency (*R.E.*) of the modified 1-ODC over the original 1-ODC, using the datasets that have been used in the illustrations in Figures 1, 2 and 3.

Table 1. Relative efficiency of the modified 1-ODC in detecting single outlier in three datasets

Data set	n	Outlier Obs.	$U(\mathbf{x}_\epsilon,\bar{\mathbf{x}}_{(\epsilon)},\mathbf{S}_{(\epsilon)})$	$tr\,\mathbf{A}_{I_1}\mathbf{S}_{(\epsilon)}^{-1}$	*R.E.*
Iris Setosa	50	42	0.3524	0.3454	1.4 : 1
Milk Transport-Cost	36	9	1.1075	1.0768	2.2 : 1
U.S Food Price	23	10	1.3110	1.2540	2.5 : 1

From the table, we see that in the Iris Setosa data the modified method isolates the outlier from the centre $\bar{\mathbf{x}}_{(\epsilon)}$ of the remaining $n-1$ observations by a distance of about one and half times the distance of the outlier from $\bar{\mathbf{x}}$ under the original method. In the Milk Transportation-cost data, the separation achieved under the modified method is about two and a quarter times that achieved under the original method. In the U.S. Food Price data, the outlier is at a distance from $\bar{\mathbf{x}}_{(\epsilon)}$ about two and half times its distance from $\bar{\mathbf{x}}$.

The relative efficiency ratios in the table obtained from the respective datasets (displayed in Figures 1, 2 and 3) reflect the extent to which the distance of $\mathbf{x}_\epsilon$ from the observation next to it (on the left) is greater with the modified method than the original method. Thus, the ratios reflect the extent of "discordancy" or significance of extremeness of the outlier in the data.

With reference to the three datasets, we can say that observation 10 is more extreme in the Food Price data than observation 9 is in the Milk Transportation-cost data, which in turn is more extreme than observation 42 is in the Iris Setosa data. If we assume the multivariate normality for these samples, observation 42 is not a discordant outlier at 5 percent level of significance since the corresponding

statistic

$$D_{42} = (n-1)U(\mathbf{x}_{42}, \bar{\mathbf{x}}, \mathbf{S}) = 12.35, \quad n = 50$$

is less than the tabulated value of 15.89 (see Table XXXII in [4]). However, by a similar test procedure, observation 9 in the Milk Transportation data and observation 10 in the Food Price data are discordant outliers. These findings are confirmed to a large extent by the relative efficiency values in the table and the plots in the figures.

Thus, in general, the value of the $R.E.$ of the modified method given in Equation (21) is an indication of the significance of extremeness of the outlier in a given dataset.

6. Conclusion

The paper considered a modification of the One-outlier Displaying Component, a projection method that is used to highlight a single outlier in multivariate data. The original method is based on the sample mean vector and the sum of squares and cross-product matrix, both of which involve the very outlier that is detected. This is a source of weakness in the method. In an attempt to eliminate this weakness, we derived a method that excludes the outlier from the mean vector and the sum of squares and cross-product matrix. The resulting method now has a number of advantages: it increases the separation of the outlier from the rest of the data points so that it appears more distinct. It also increases the general dispersion in the projected data so that the presence of multiple outliers could be highlighted.

The paper also derived a measure of relative efficiency of the modified method over the original method. This measure gives an indication of the extent of extremeness of the outlier in a given dataset.

Appendix: U.S. Food Price Data

No.	City	Bread	Burger	Milk	Oranges	Tomates
1	Atlanta	24.5	94.5	73.9	80.1	41.6
2	Baltimore	26.5	91.0	67.5	74.6	53.3
3	Boston	29.7	100.8	61.4	104.0	59.6
4	Buffalo	22.8	86.6	65.3	118.4	51.2
5	Chicago	26.7	86.7	62.7	105.9	51.2
6	Cincinnati	25.3	102.5	63.3	99.3	45.6
7	Cleveland	22.8	88.8	52.4	110.9	46.8
8	Dallas	23.3	85.5	62.5	117.9	41.8
9	Detroit	24.1	93.7	51.5	109.7	52.4
10	Honolulu	29.3	105.9	80.2	133.2	61.7
11	Houston	22.3	83.6	67.8	108.6	42.4
12	Kansas city	26.1	88.9	65.4	100.9	43.2
13	Los Angeles	26.9	89.3	56.2	82.7	38.4

Contd.

No.	City	Bread	Burger	Milk	Oranges	Tomates
14	Milwaukee	20.3	89.6	53.8	111.8	53.9
15	Minneapolis	24.6	92.2	51.9	106.0	50.7
16	New York	30.8	110.7	66.0	107.3	62.6
17	Philadelphia	24.5	92.3	66.7	98.0	61.7
18	Pittsburgh	26.2	95.4	60.2	117.1	49.3
19	St. Louis	26.5	92.4	60.8	115.1	46.2
20	San Diego	25.5	83.7	57.0	92.8	35.4
21	San Francisco	26.3	87.1	58.3	101.8	41.5
22	Seattle	22.5	77.7	62.0	91.1	44.9
23	Washington, DC	24.2	93.8	66.0	81.6	46.2

Source: Estimated Retail Food Prices by Cities, March 1973,
U.S. Department of Labour, Bureau of Labour Statistics [11]

References

[1] E. Anderson (1939), The Irises of the Gaspe Peninsula, *Bulletin of the American Iris Society* **59**, 2–5.

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

[3] J. Bacon-Shone and W.K. Fung (1987), A new graphical method for detecting single and multiple ouliers in univariate and multivariate data, *Applied Statistics* **36**(2), 153–162.

[4] V. Barnett and T. Lewis (1994), *Outliers in Statistical Data*, 3rd edition, Wiley and Sons, New York.

[5] G. Dahlquist and A. Bjorck (1974), *Numerical Methods*, Prentice-Hall, Englewood Cliffs, NJ.

[6] N.R.J. Fieller (1976), *Some Problems Related to the Rejection of Outlying Observations*, Ph.D. Thesis, University of Hull.

[7] B.K. Gordor (1994), *Some Informal Methods for the Detection and Display of Outliers in Data*, Ph.D. Thesis, University of Sheffield.

[8] K.S. Miller (1981), On the inverse of the sum of matrices, *JSTOR* **54**(2), 67–72.

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

[10] W.H. Press, S.A. Teukolsky, W.T. Vetterling and B.P. Flannery (2007), *Numerical Recipes: The Art of Scientific Computing*, 3rd edition, Cambridge University Press, New York.

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

B.K. Nkansah, *Department of Mathematics and Statistics, University of Cape Coast, Ghana*.
E-mail: gyaabeng@yahoo.co.uk

B.K. Gordor, *Department of Mathematics and Statistics, University of Cape Coast, Ghana*.
E-mail: benkgordor@yahoo.co.uk

Received November 11, 2011
Accepted May 23, 2012