American Journal of Theoretical and Applied Statistics
2013; 2(2) : 29-37
Published online April 2, 2013 (http://www.sciencepublishinggroup.com/j/ajtas)
doi: 10.11648/j.ajtas.20130202.14

Discordancy in reduced dimensions of outliers in high-dimensional datasets: application of an updating formula

B. K. Nkansah, B. K. Gordor

Department of Mathematics and Statistics, Cape Coast, Ghana

Email address:
gyaabeng@yahoo.co.uk (B. K. Nkansah), benkgordor@yahoo.co.uk (B. K. Gordor)

To cite this article:
B. K. Nkansah, B. K. Gordor. Discordancy in Reduced Dimensions of Outliers in High-Dimensional Datasets: Application of an Updating Formula, *American Journal of Theoretical and Applied Statistics,* Vol. 2, No. 2, 2013, pp. 29-37. doi: 10.11648/j.ajtas.20130202.14

Abstract: In multivariate outlier studies, the sum of squares and cross-product (SSCP) is an important property of the data matrix. For example, the much used Mahalanobis distance and the Wilk's ratio make use of SSCP matrices. One of the SSCP matrices involved in outlier studies is the matrix for the set of multiple outliers in the data. In this paper, an explicit expression for this matrix is derived. It has then been shown that in general the discordancy of multiple outliers is preserved along Multiple-Outlier Displaying Components with much lower dimensions than the original high-dimensional dataset.

Keywords: Outlier Detection, Discordancy, Updating Formula, Outlier Displaying Components

1. Introduction

The SSCP matrices have important uses in multivariate data analysis. For example, the well known Wilk's *k*-outlier ratio given by

$$r_k = \frac{|\mathbf{S}_{(I_k)}|}{|\mathbf{S}|} \qquad (1.1)$$

involve two matrices which are the SSCP of the entire sample S, and the SSCP for the remaining sample, $\mathbf{S}_{I_k}$ when a set of k outliers are deleted. A third matrix is the SSCP of the k-tuple outliers, $\mathbf{A}_{I_k}$. The three matrices are related by the equation

$$\mathbf{S}_{(I_k)} = \mathbf{S} - \mathbf{A}_{I_k} \qquad (1.2)$$

We know that for a p-dimensional random sample $\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_n$, S is given by

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

Suppose I_k is an indexed set of k outliers from the sample. Then the SSCP, $\mathbf{S}_{(I_k)}$ for the remaining $(n-k)$ observations, $\mathbf{x}_j$, $j \notin I_k$, is given by

$$\mathbf{S}_{(I_k)} = \sum_{j \notin I}^{n} (\mathbf{x}_j - \bar{\mathbf{x}}_{(I_k)})(\mathbf{x}_j - \bar{\mathbf{x}}_{(I_k)})' \text{ where } \bar{\mathbf{x}}_{(I_k)} = \frac{1}{n-k}\sum_{j \notin I}^{n} \mathbf{x}_j .$$

Equation (1.2) is usually referred to as an updating formula for $\mathbf{S}_{(I_k)}$.

It appears that in the multiple outlier case, an expression for $\mathbf{A}_{I_k}$ similar to those for S and $\mathbf{S}_{(I_k)}$ has not been given the needed attention. That is, direct use for the matrix $\mathbf{A}_{I_k}$ seems unpopular. In the single outlier case, $\mathbf{A}_{I_k}$ is given as

$$\mathbf{A}_{I_1} = \frac{n}{n-1}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})', \quad i = 1, 2, \cdots, n \qquad (1.3)$$

and substituting into Equation (1.2) gives the corresponding updating formula [2]. Caroni and Prescott [3] make a rather consecutive use of the expression in Equation (1.1) for the most extreme observation after the previous most extreme one is deleted. This is equivalent to a consecutive use of the expression

$$D_1 = 1 - \frac{n}{n-1}(\mathbf{x}_j - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_j - \bar{\mathbf{x}}) \qquad (1.4)$$

(see Remark 2.1 for proof). In this case, the k-tuple of extreme observations are not examined simultaneously. The objective of the next section is to obtain a more `simulta-

neous' and explicit expression for $\mathbf{A}_{I_k}$. The result would enable us to concentrate on the k extreme observations each time instead of the remaining $(n-k)$. This will be used in Section 3 to generalize a result that enables us to obtain a reduced-dimensional dataset, equivalent of the original high-dimensional dataset, in which the discordancy of the k-tuple outlier is preserved. Our approach to discordancy of outliers follows that of Wilks [12]. Definitions and tests for discordancy of multivariate outliers are well presented in [2]. The last two sections provide illustration of the concept developed in the paper and conclusions of the results.

2. Derivation of an Updating Formula

Suppose that in the random matrix $\mathbf{X}_{n\times p}=[\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, \cdots, \mathbf{x}_n]'$ of p-dimensional measurements, k of the observations belong to the indexed set $I_k=\{i_1, i_2, \cdots, i_k\}$ of labelled outliers. Define an $n\times n$ matrix $\mathbf{D}_n$ by

$$\mathbf{D}_n=\mathbf{I}_n-\frac{1}{n}\mathbf{1}_n\mathbf{1}_n', \tag{2.1}$$

where $\mathbf{1}_n=(1, 1, \cdots, 1)'\in\Re^n$ (See e.g. [5]; [9]). The matrices S and $\mathbf{S}_{I_k}$ may be written as $\mathbf{S}=\mathbf{x}'\mathbf{D}_n\mathbf{x}$ and $\mathbf{S}_{(I_k)}=\mathbf{x}'\mathbf{C}_1\mathbf{x}$, where $\mathbf{C}_1=\mathrm{diag}(\mathbf{D}_{n-k}, \mathbf{0})$. From Equation (2.1)

$$\begin{aligned}\mathbf{A}_{I_k}&=\mathbf{x}'\mathbf{D}_n\mathbf{x}-\mathbf{x}'\mathbf{C}_1\mathbf{x}\\&=\mathbf{x}'(\mathbf{D}_n-\mathbf{C}_1)\mathbf{x}\end{aligned}$$

If $\mathbf{C}_2=\mathbf{D}_n-\mathbf{C}_1$, then

$$\mathbf{A}_{I_k}=\mathbf{x}'\mathbf{C}_2\mathbf{x} \tag{2.2}$$

Now, partitioning the matrix $\mathbf{D}_n$ in Equation (2.1) as

$$\mathbf{D}_n=\left(\begin{array}{c|c}\mathbf{I}_{n-k}-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}' & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_k'\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}_{n-k}' & \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}_k'\end{array}\right)$$

and $\mathbf{C}_1$ as

$$\mathbf{C}_1=\left(\begin{array}{c|c}\mathbf{I}_{n-k}-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}' & \mathbf{0}_{(n-k)\times k}\\ \hline \mathbf{0}_{k\times(n-k)} & \mathbf{0}_{k\times k}\end{array}\right)$$

Thus, $\mathbf{C}_2$ is given by

$$\begin{aligned}\mathbf{C}_2&=\mathbf{D}_n-\mathbf{C}_1\\&=\left(\begin{array}{c|c}\mathbf{I}_{n-k}-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}' & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_k'\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}_{n-k}' & \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}_k'\end{array}\right)-\left(\begin{array}{c|c}\mathbf{I}_{n-k}-\frac{1}{n-k}\mathbf{1}_{n-k}\mathbf{1}_{n-k}' & \mathbf{0}_{(n-k)\times k}\\ \hline \mathbf{0}_{k\times(n-k)} & \mathbf{0}_{k\times k}\end{array}\right)\\&=\left(\begin{array}{c|c}\left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}_{n-k}' & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_k'\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}_{n-k}' & \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}_k'\end{array}\right)\end{aligned} \tag{2.3}$$

The matrix $\mathbf{C}_2$ may be written in another useful way as a sum of two matrices which is

$$\begin{aligned}\mathbf{C}_2&=-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}'+\left(\begin{array}{c|c}\frac{1}{n-k}\mathbf{1}_{n-k}\mathbf{1}_{n-k}' & \mathbf{0}_{(n-k)\times k}\\ \hline \mathbf{0}_{k\times(n-k)} & \mathbf{I}_k\end{array}\right)\\&=-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}_{n-k}'+\mathrm{diag}\left[-\frac{1}{n-k}\mathbf{1}_{n-k}\mathbf{1}_{n-k}', \quad \mathbf{I}_k\right]\end{aligned}$$

The matrix $\mathbf{C}_2$ in Equation (2.3) is of the form

$$\mathbf{C}_2=\left(\begin{array}{c|c}\mathbf{C}_2^{11} & \mathbf{C}_2^{12}\\ \hline \mathbf{C}_2^{21} & \mathbf{C}_2^{22}\end{array}\right)$$

where $\mathbf{C}_2^{11}$ is of dimension $(n-k)\times(n-k)$; $\mathbf{C}_2^{12}$ is of dimension $(n-k)\times k$; $\mathbf{C}_2^{21}$ is of dimension $k\times(n-k)$; $\mathbf{C}_2^{22}$ is of dimension $k\times k$.

In order to make it comparable to the format of $\mathbf{C}_2$, $\mathbf{X}'_{p\times n}$ is partitioned as

$$\mathbf{X}'_{p\times n}=\left(\begin{array}{c|c}\mathbf{x}'_{(I)(n-k)\times(n-k)} & \mathbf{x}'_{I(n-k)\times k}\\ \hline \mathbf{x}'_{(I)\{p-(n-k)\}\times(n-k)} & \mathbf{x}'_{I\{p-(n-k)\}\times k}\end{array}\right)$$

That is, the last k rows of the data matrix $\mathbf{X}_{n\times p}$ have been labelled as the outlier observations, denoted by a $k\times p$ matrix $\mathbf{x}_I$ and $\mathbf{x}_{(I)}$ is $(n-k)\times p$ matrix without the outliers.

Let the general element of the product matrix $\mathbf{X}'_{p\times n}\mathbf{C}_2$ be $\mathbf{H}_{ij}$. Then $\mathbf{H}_{11}$, $\mathbf{H}_{12}$, $\mathbf{H}_{21}$ and $\mathbf{H}_{22}$ are derived as follows.

$$\begin{aligned}\mathbf{H}_{11}&=\frac{k}{n(n-k)}\mathbf{x}'_{(I)(n-k)\times(n-k)}\mathbf{1}_{n-k}\mathbf{1}_{n-k}'-\frac{1}{n}\mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\mathbf{1}_{n-k}'\\&=\left(\frac{k}{n(n-k)}\mathbf{x}'_{(I)(n-k)\times(n-k)}\mathbf{1}_{n-k}-\frac{1}{n}\mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\right)\mathbf{1}_{n-k}'\\&=\left\{\frac{nk}{n(n-k)}\left(\frac{1}{n}\mathbf{x}_g'^{(1)}\mathbf{1}_n-\frac{1}{n}\mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\right)-\frac{1}{n}\mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\right\}\mathbf{1}_{n-k}'\\&=\left(\frac{k}{n-k}\bar{\mathbf{x}}_g'^{(1)}-\frac{1}{n-k}\mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\right)\mathbf{1}_{n-k}'\\&=\frac{1}{n-k}\left(\bar{\mathbf{x}}_g'^{(1)}\mathbf{1}_k'\mathbf{1}_k-\mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\right)\mathbf{1}_{n-k}'\\&=\frac{1}{n-k}\left(\bar{\mathbf{x}}_g'^{(1)}\mathbf{1}_k'-\mathbf{x}'_{I(n-k)\times k}\right)\mathbf{1}_k\mathbf{1}_{n-k}'\end{aligned}$$

where $\mathbf{x}_g'^{(1)}$ is data on the first g dimensions, $g<p$, and $\bar{\mathbf{x}}_g'^{(1)}$ is the first g components of the mean vector. The $\mathbf{H}_{12}$ element is

$$\begin{aligned}\mathbf{H}_{12} &= -\frac{1}{n}\left(\mathbf{x}'_{(I)(n-k)\times(n-k)}\mathbf{1}_{n-k}\mathbf{1}'_k + \mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\mathbf{1}'_k\right) + \mathbf{x}'_{I(n-k)\times k} \\ &= -\frac{1}{n}\left(\mathbf{x}'_{(I)(n-k)\times(n-k)}\mathbf{1}_{n-k} + \mathbf{x}'_{I(n-k)\times k}\mathbf{1}_k\right)\mathbf{1}'_k + \mathbf{x}'_{I(n-k)\times k} \\ &= -\overline{\mathbf{x}}'^{(1)}_g\mathbf{1}'_k + \mathbf{x}'_{I(n-k)\times k}\end{aligned}$$

The $\mathbf{H}_{21}$ element is

$$\begin{aligned}\mathbf{H}_{21} &= \left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{x}'_{(I)\{p-(n-k)\}\times(n-k)}\mathbf{1}_{n-k}\mathbf{1}'_{n-k} - \frac{1}{n}\mathbf{x}'_{I\{p-(n-k)\}\times k}\mathbf{1}_k\mathbf{1}'_{n-k} \\ &= \left(\frac{k}{n(n-k)}\mathbf{x}'_{(I)\{p-(n-k)\}\times(n-k)}\mathbf{1}_{n-k} - \frac{1}{n}\mathbf{x}'_{I\{p-(n-k)\}\times k}\mathbf{1}_k\right)\mathbf{1}'_{n-k} \\ &= \left\{\frac{nk}{n(n-k)}\left(\frac{1}{n}\mathbf{x}'^{(2)}_{(p-g)}\mathbf{1}_n - \frac{1}{n}\mathbf{x}'_{I\{p-(n-k)\}\times k}\mathbf{1}_k\right) - \frac{1}{n}\mathbf{x}'_{I\{p-(n-k)\}\times k}\mathbf{1}_k\right\}\mathbf{1}'_{n-k} \\ &= \left(\frac{k}{n-k}\overline{\mathbf{x}}'^{(2)}_{(p-g)} - \frac{1}{n-k}\mathbf{x}'_{I\{p-(n-k)\}\times k}\mathbf{1}_k\right)\mathbf{1}'_{n-k} \\ &= \frac{1}{n-k}\left(\overline{\mathbf{x}}'^{(2)}_{(p-g)}\mathbf{1}'_k\mathbf{1}_k - \mathbf{x}'_{I\{p-(n-k)\}\times k}\mathbf{1}_k\right)\mathbf{1}'_{n-k} \\ &= \frac{1}{n-k}\left(\overline{\mathbf{x}}'^{(2)}_{(p-g)}\mathbf{1}'_k - \mathbf{x}'_{I\{p-(n-k)\}\times k}\right)\mathbf{1}_k\mathbf{1}'_{n-k}\end{aligned}$$

where $\mathbf{x}'^{(2)}_{(p-g)}$ is data on the last $(p-g)$ dimensions, and $\overline{\mathbf{x}}'^{(2)}_{(p-g)}$ is the second $(p-g)$ components of the mean vector.

The last element of $\mathbf{X}'_{p\times n}\mathbf{C}_2$ is given by

$$\begin{aligned}\mathbf{H}_{22} &= -\frac{1}{\mathrm{n}}\mathbf{x}'_{(\mathrm{I})\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times(\mathrm{n}-\mathrm{k})}\mathbf{1}_{\mathrm{n}-\mathrm{k}}\mathbf{1}'_{\mathrm{k}} \\ &+ \mathbf{x}'_{\mathrm{I}\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times\mathrm{k}} - \frac{1}{\mathrm{n}}\mathbf{x}'_{\mathrm{I}\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times\mathrm{k}}\mathbf{1}_{\mathrm{k}}\mathbf{1}'_{\mathrm{k}} \\ &= -\frac{1}{\mathrm{n}}\begin{pmatrix}\mathbf{x}'_{(\mathrm{I})(\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times(\mathrm{n}-\mathrm{k})}\mathbf{1}_{\mathrm{n}-\mathrm{k}} \\ +\ \mathbf{x}'_{\mathrm{I}\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times\mathrm{k}}\mathbf{1}_{\mathrm{k}}\end{pmatrix}\mathbf{1}'_{\mathrm{k}} + \mathbf{x}'_{\mathrm{I}\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times\mathrm{k}} \\ &= -\overline{\mathbf{x}}'^{(2)}_{(\mathrm{p}-\mathrm{g})}\mathbf{1}'_{\mathrm{k}} + \mathbf{x}'_{\mathrm{I}\{\mathrm{p}-(\mathrm{n}-\mathrm{k})\}\times\mathrm{k}}\end{aligned}$$

Therefore,

$$\mathbf{X}'_{p\times n}\mathbf{C}_2 = \left(\begin{array}{c|c}\frac{1}{n-k}\left(\overline{\mathbf{x}}^{(1)}_g\mathbf{1}'_k - \mathbf{x}'_{I(n-k)\times k}\right)\mathbf{1}_k\mathbf{1}'_{n-k} & -\overline{\mathbf{x}}^{(1)}_g\mathbf{1}'_k + \mathbf{x}'_{I(n-k)\times k} \\ \frac{1}{n-k}\left(\overline{\mathbf{x}}'^{(2)}_{(p-g)}\mathbf{1}'_k - \mathbf{x}'_{I\{p-(n-k)\}\times k}\right)\mathbf{1}_k\mathbf{1}'_{n-k} & -\overline{\mathbf{x}}'^{(2)}_{(p-g)}\mathbf{1}'_k + \mathbf{x}'_{I\{p-(n-k)\}\times k}\end{array}\right)$$

In this matrix, for the ith row of the first $p\times(n-k)$ part, all of the $(n-k)$ columns contain the same element which is of the form

$$\frac{1}{n-k}\sum_{j\in I}(\overline{\mathbf{x}}_i - \mathbf{x}_{ji}), \qquad i = 1,\ 2,\ \cdots,\ p$$

The $(i,\ j)$ element of the second $p\times k$ part is of the form $-(\overline{\mathbf{x}}_i - \mathbf{x}_{ji})$. Thus, $\mathbf{X}'_{p\times n}\mathbf{C}_2$ simplifies as

$$\mathbf{X}'\mathbf{C}_2 = \left\{\frac{1}{n-k}\left(\overline{\mathbf{x}}\mathbf{1}'_k - \mathbf{x}'_{I_k}\right)\mathbf{1}_k\mathbf{1}'_{n-k} \;\middle|\; -\overline{\mathbf{x}}\mathbf{1}'_k + \mathbf{x}'_{I_k}\right\}.$$

Now partition the data matrix, $\mathbf{X}_{n\times p}$ in a manner similar to the above as

$$\mathbf{X}_{n\times p} = \left(\mathbf{x}_{(I)(n-k)\times p} \;\middle|\; \mathbf{x}_{Ik\times p}\right)'$$

Hence,

$$\begin{aligned}\mathbf{A}_{I_k} &= \left\{\frac{1}{n-k}\left(\overline{\mathbf{x}}\mathbf{1}'_k - \mathbf{x}'_{I_k}\right)\mathbf{1}_k\mathbf{1}'_{n-k} \;\middle|\; -\overline{\mathbf{x}}\mathbf{1}'_k + \mathbf{x}'_{I_k}\right\}\left(\mathbf{x}_{(I)} \;\middle|\; \mathbf{x}_{I_k}\right)' \\ &= \left\{\frac{1}{n-k}\left(\overline{\mathbf{x}}\mathbf{1}'_k - \mathbf{x}'_{I_k}\right)\mathbf{1}_k\mathbf{1}'_{n-k}\right\}\left(\mathbf{x}'_{(I)}\mathbf{1}_{n-k}\right) + \left(-\overline{\mathbf{x}}\mathbf{1}'_k + \mathbf{x}'_{I_k}\right)\mathbf{x}_{I_k}\end{aligned}$$

Further simplification gives $(i,\ t)$ element of $\mathbf{A}_{I_k}$ as

$$\mathbf{A}_{I_k} = \left(\frac{1}{n-k}\sum_{j\in I}\left(\overline{\mathbf{x}}_i - \mathbf{x}_{ji}\right)\left(\sum_{j\notin I}\mathbf{x}_{jt}\right) - \sum_{j\in I}\left(\overline{\mathbf{x}}_i - \mathbf{x}_{ji}\right)\mathbf{x}_{jt}\right) \quad (2.4)$$

We substitute

$$n\left\{\frac{1}{n}\left(\sum_{j\notin I}\mathbf{x}_{jt} + \sum_{j\in I}\mathbf{x}_{jt}\right) - \frac{1}{n}\sum_{j\in I}\mathbf{x}_{jt}\right\}$$

for $\sum_{j\notin I}\mathbf{x}_{jt}$ and note that the first term involving the two summands in the inner brackets is simply $\frac{1}{n}\sum_{j=1}^{n}\mathbf{x}_{jt} = \overline{\mathbf{x}}_t$, which is the tth component of the sample mean vector, $\overline{\mathbf{x}}$. Equation (2.4) then becomes

$$\mathbf{A}_{\mathrm{I}_\mathrm{k}} = \begin{pmatrix}\frac{\mathrm{n}}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\overline{\mathbf{x}}_\mathrm{i} - \mathbf{x}_{\mathrm{ji}}\right)\overline{\mathbf{x}}_\mathrm{t} \\ -\frac{1}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\overline{\mathbf{x}}_\mathrm{i} - \mathbf{x}_{\mathrm{ji}}\right)\left(\sum_{\mathrm{j}\in\mathrm{I}}\mathbf{x}_{\mathrm{jt}}\right) - \sum_{\mathrm{j}\in\mathrm{I}}\left(\overline{\mathbf{x}}_\mathrm{i} - \mathbf{x}_{\mathrm{ji}}\right)\mathbf{x}_{\mathrm{jt}}\end{pmatrix}$$

Again, observe that $\sum_{j\in I}\mathbf{x}_{jt} = k\overline{\mathbf{x}}_{I_k}$ and then further simplification gives the $(i,\ t)$ element of $\mathbf{A}_{I_k}$ as

$$\mathbf{A}_{\mathrm{I}_\mathrm{k}} = \begin{pmatrix}\frac{\mathrm{n}}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\overline{\mathbf{x}}_\mathrm{i} - \mathbf{x}_{\mathrm{ji}}\right)\left(\overline{\mathbf{x}}_\mathrm{i} - \mathbf{x}_{\mathrm{jt}}\right)' \\ -\frac{\mathrm{k}}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\overline{\mathbf{x}}_\mathrm{i} - \mathbf{x}_{\mathrm{ji}}\right)\left(\overline{\mathbf{x}}_{\mathrm{I}_\mathrm{k}} - \mathbf{x}_{\mathrm{jt}}\right)'\end{pmatrix} \quad (2.5)$$

Finally, we drop i and t in Equation (2.5), and write $\mathbf{A}_{I_k}$ as the difference of the two $p\times p$ matrices as

$$\mathbf{A}_{\mathrm{I}_\mathrm{k}} = \begin{pmatrix}\frac{\mathrm{n}}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_\mathrm{j} - \overline{\mathbf{x}}\right)\left(\mathbf{x}_\mathrm{j} - \overline{\mathbf{x}}\right)' \\ -\frac{\mathrm{k}}{\mathrm{n}-\mathrm{k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_\mathrm{j} - \overline{\mathbf{x}}\right)\left(\mathbf{x}_\mathrm{j} - \overline{\mathbf{x}}_{\mathrm{I}_\mathrm{k}}\right)'\end{pmatrix} \quad (2.6)$$

This is the general expression for $\mathbf{A}_{I_k}$ which satisfies Equation (1.2) and is referred to as the updating formula.

2.1. Some Remarks on the Updating Formula

Relationship between Wilk's One-Outlier Ratio and Generalized Distance

For the single outlier case, suppose the index set is $I_k = \{i\}$ $(i = 1, 2, \cdots, n\}$. Then $\bar{\mathbf{x}}_{I_k} = \mathbf{x}_i$, and from Equation (2.6), $\mathbf{A}_{I_1}$ for the ith observation is given as

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

which is equal to Equation (1.3). In particular, if $I_k = \{n\}$, then from Equation (1.2), we obtain

$$\begin{aligned}\mathbf{S}_{(I_k)} &= \mathbf{S} - \mathbf{A}_{I_1} \\ &= \mathbf{S} - \frac{n}{n-1}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'\end{aligned}$$

multiplying both sides by $\mathbf{S}^{-1}$ and taking determinant of both sides, we obtain

$$\frac{|\mathbf{S}_{(I_k)}|}{|\mathbf{S}|} = \left|\mathbf{I}_n - \frac{n}{n-1}\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'\right| \quad (2.7)$$

Observe that the left hand side of Equation (2.7) is r_k, the Wilk's ratio. To determine the determinant on the right hand side, we note that the $p \times p$ matrix $(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'$ is of rank 1. Thus, $\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'$ is also of rank 1, and hence all except one of its eigenvalues is zero (see, e.g., [6]; [10]). That single non-zero eigenvalue may be written as

$$\operatorname{tr}\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})' = (\mathbf{x}_n - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})$$

It follows that

$$1 - \frac{n}{n-1}\operatorname{tr}\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})'$$

is an eigenvalue of the matrix on the right hand side of Equation (2.7). The remaining $(p-1)$ eigenvalues are all equal to 1. The determinant of the matrix on the right hand side, which is the product of its eigenvalues, is therefore equal to

$$1 - \frac{n}{n-1}(\mathbf{x}_n - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})$$

This expression is equal to the right hand side of Equation (1.4) which is r_k for one outlier. We note that $(\mathbf{x}_n - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})$ is the Mahalanobis generalized distance, $U_{(n)}$ of $\mathbf{x}_n$ from $\bar{\mathbf{x}}$ which in this case is the largest of all $U_{(i)}$. Therefore, for $k = 1$, r_k is a direct function of $U_{(n)}$. This result shows that for detecting one outlier, both the Mahalanobis distance and the Wilk's ratio produces the same result.

3. Discordancy of Multiple Outliers

In this section, we will apply the expression for the SSCP matrix, $\mathbf{A}_{I_k}$ for a set of k-tuple extreme observations in a dataset. It will be shown that the value of the Wilks' k-outlier scatter ratio obtained from the original high-dimensional dataset is the same as that obtained from a reduced-dimensional dataset, equivalent of the original.

Let $\mathbf{P}_{p\times k}$ denote the matrix of eigenvectors corresponding to the k non-zero eigenvalues of the matrix

$$\mathbf{E}_k = \mathbf{S}^{-1}\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)$$

where the indexed set is as defined. The eigenvectors corresponding to the k non-zero eigenvalues of $\mathbf{E}_k$ will generally be referred to as the k-Outlier Displaying Components, or k-ODC ([7]; [4]).

Let $\mathbf{V}_k = [\mathbf{V}(:,1) \quad \mathbf{V}(:,2) \quad \mathbf{V}(:,3), \cdots, \quad \mathbf{V}(:,k)]$ be the first k columns of the matrix $\mathbf{V}$ of eigenvectors associated with the eigenvalues of $\mathbf{E}_k$. Then we have

$$\left[\mathbf{S}^{-1}\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right) - \Lambda\mathbf{I}\right]\mathbf{V} = 0$$

Consider the projection $\mathbf{Y} = \mathbf{X}_{n\times p} * \mathbf{V}_{p\times p}$. Then the variance-covariance matrix of Y is $\mathbf{S}_\mathbf{Y} = \mathbf{V}'\mathbf{S}\mathbf{V}$. This SSCP matrix is generally of the form

$$\mathbf{V}'\mathbf{S}\mathbf{V} = \left(\begin{array}{c|c} \mathbf{M}_{k\times k} & \mathbf{0}_{k\times(p-k)} \\ \hline \mathbf{0}'_{(p-k)\times k} & \mathbf{N}_{(p-k)\times(p-k)} \end{array}\right) \quad (3.1)$$

where $\mathbf{M}_{k\times k} = \operatorname{diag}(\kappa_1, \kappa_2, \cdots, \kappa_k)$ with some positive numbers κ_i, $i = 1, 2, \cdots, k$ and N is a non-diagonal matrix. Now divide the jth column of V by the square root of the corresponding jth diagonal element of $\mathbf{V}'\mathbf{S}\mathbf{V}$. By this we transform V into a matrix U such that

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \left(\begin{array}{c|c} \mathbf{I}_{k\times k} & \mathbf{0}_{k\times(p-k)} \\ \hline \mathbf{0}'_{(p-k)\times k} & \mathbf{G}_{(p-k)\times(p-k)} \end{array}\right) \quad (3.2)$$

where G is a non-diagonal matrix with diagonal elements equal to 1. If U is partitioned as

$$\mathbf{U} = \left(\mathbf{P}_{p\times k} \;\middle|\; \mathbf{Q}_{p\times(p-k)}\right),$$

then

$$\begin{aligned}\mathbf{P}'\mathbf{S}\mathbf{P} &= \mathbf{I}_k \\ \mathbf{Q}'\mathbf{S}\mathbf{Q} &= \mathbf{G}_{p-k} \\ \mathbf{U}'\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)\mathbf{U} &= \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k, 0, \cdots, 0) \\ &= \Lambda_p \\ \mathbf{P}'\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)\mathbf{P} &= \operatorname{diag}(\lambda_1, \lambda_2, \cdots, \lambda_k) \\ \mathbf{Q}'\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)\mathbf{Q}' &= \mathbf{0}_{p-k}\end{aligned}$$

where $\lambda_1, \lambda_2, \cdots, \lambda_k$ are the k non-zero eigenvalues of $\mathbf{E}_k$.

Define a new set of variables, $\mathbf{Z}_{k\times n}$ by the transformation $\mathbf{Z}=\mathbf{P}'\mathbf{X}$. By definition, the Wilks' k-tuple outlier test statistic using $\mathbf{Z}_{k\times n}$ is

$$r_k^{(\mathbf{Z})}=\frac{\left|\mathbf{S}_{(I_k)}^{(\mathbf{Z})}\right|}{\left|\mathbf{S}^{(\mathbf{Z})}\right|}$$

where $\mathbf{S}^{(\mathbf{Z})}=\mathbf{Z}\mathbf{Z}'$ and $\mathbf{S}_{(I_k)}^{(\mathbf{Z})}$ is the corresponding matrix with observations in the indexed set I_k removed from the sample. Substituting for $\mathbf{Z}$ we obtain

$$r_k^{(\mathbf{Z})}=\frac{\left|\mathbf{P}'\mathbf{S}_{(I_k)}\mathbf{P}\right|}{\left|\mathbf{P}'\mathbf{S}\mathbf{P}\right|}=\left|\mathbf{P}'\mathbf{S}_{(I_k)}\mathbf{P}\right| \quad (3.3)$$

Now from Section 2, we substitute for $\mathbf{A}_{I_k}$ in Equation (1.2) to obtain

$$\mathbf{S}_{(\mathrm{I_k})}=\mathbf{S}-\frac{\mathrm{n}}{\mathrm{n-k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)'+\frac{\mathrm{k}}{\mathrm{n-k}}\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'$$

Thus,

$$\begin{aligned}\mathbf{U}'\mathbf{S}_{(\mathrm{I_k})}\mathbf{U}&=\mathbf{U}'\mathbf{S}\mathbf{U}-\frac{\mathrm{n}}{\mathrm{n-k}}\mathbf{U}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)'\right]\mathbf{U}\\&+\frac{\mathrm{k}}{\mathrm{n-k}}\mathbf{U}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'\right]\mathbf{U}\\&=\mathbf{U}'\mathbf{S}\mathbf{U}-\frac{\mathrm{n}}{\mathrm{n-k}}\Lambda_{\mathrm{p}}+\frac{\mathrm{k}}{\mathrm{n-k}}\mathbf{U}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'\right]\mathbf{U}\\\mathbf{P}'\mathbf{S}_{(\mathrm{I_k})}\mathbf{P}&=\mathbf{P}'\mathbf{S}\mathbf{P}-\frac{\mathrm{n}}{\mathrm{n-k}}\mathbf{P}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)'\right]\mathbf{P}\\&+\frac{\mathrm{k}}{\mathrm{n-k}}\mathbf{P}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'\right]\mathbf{P}\\&=\mathbf{I}_{\mathrm{k}}-\frac{\mathrm{n}}{\mathrm{n-k}}\Lambda_{\mathrm{k}}+\frac{\mathrm{k}}{\mathrm{n-k}}\mathbf{P}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'\right]\mathbf{P}\end{aligned}$$

We note that

$$\mathbf{Q}'\left[\sum_{j\in I}\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)'\right]\mathbf{Q}=\mathbf{0},$$

$$\mathbf{Q}'\left[\sum_{j\in I}\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)'\right]\mathbf{P}=\mathbf{0},$$

$$\mathbf{Q}'\left[\sum_{j\in I}\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}_{I_k}\right)'\right]\mathbf{Q}=\mathbf{0},$$

and

$$\mathbf{P}'\left[\sum_{j\in I}\left(\mathbf{x}_j-\bar{\mathbf{x}}\right)\left(\mathbf{x}_j-\bar{\mathbf{x}}_{I_k}\right)'\right]\mathbf{Q}=\mathbf{0}$$

Hence, obtain

$$\begin{aligned}\mathbf{P}'\mathbf{S}_{(\mathrm{I_k})}\mathbf{Q}&=\mathbf{P}'\mathbf{S}\mathbf{Q}-\frac{\mathrm{n}}{\mathrm{n-k}}\mathbf{P}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)'\right]\mathbf{Q}\\&+\frac{\mathrm{k}}{\mathrm{n-k}}\mathbf{P}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'\right]\mathbf{Q}\\&=\mathbf{Q}'\mathbf{S}_{(\mathrm{I_k})}\mathbf{P}\\&=\mathbf{0}\end{aligned}$$

$$\begin{aligned}\mathbf{Q}'\mathbf{S}_{(\mathrm{I_k})}\mathbf{Q}&=\mathbf{Q}'\mathbf{S}\mathbf{Q}-\frac{\mathrm{n}}{\mathrm{n-k}}\mathbf{Q}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)'\right]\mathbf{Q}\\&+\frac{\mathrm{k}}{\mathrm{n-k}}\mathbf{Q}'\left[\sum_{\mathrm{j}\in\mathrm{I}}\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}\right)\left(\mathbf{x}_{\mathrm{j}}-\bar{\mathbf{x}}_{\mathrm{I_k}}\right)'\right]\mathbf{Q}\\&=\mathbf{Q}'\mathbf{S}\mathbf{Q}\\&=\mathbf{G}_{\mathrm{p-k}}\end{aligned}$$

Since, $\mathbf{U}=\left(\mathbf{P}_{p\times k}\mid\mathbf{Q}_{p\times(p-k)}\right)$, we obtain

$$\mathbf{U}'\mathbf{S}_{(I_k)}\mathbf{U}=\left(\begin{array}{c|c}\mathbf{P}'\mathbf{S}_{(I_k)}\mathbf{P} & \mathbf{P}'\mathbf{S}_{(I_k)}\mathbf{Q}\\\hline\mathbf{Q}'\mathbf{S}_{(I_k)}\mathbf{P} & \mathbf{Q}'\mathbf{S}_{(I_k)}\mathbf{Q}\end{array}\right) \quad (3.4)$$

Substituting the relevant results in Equation (3.4) gives

$$\mathbf{U}'\mathbf{S}_{(I_k)}\mathbf{U}=\left(\begin{array}{c|c}\mathbf{P}'\mathbf{S}_{(I_k)}\mathbf{P} & \mathbf{0}\\\hline\mathbf{0} & \mathbf{Q}'\mathbf{S}_{(I_k)}\mathbf{Q}\end{array}\right)$$

Taking determinants gives

$$\begin{aligned}\left|\mathbf{U}'\mathbf{S}_{(I_k)}\mathbf{U}\right|&=\left|\mathbf{P}'\mathbf{S}_{(I_k)}\mathbf{P}\right|\left|\mathbf{G}_{p-k}\right|\\&=\left|\mathbf{G}\right|r_k^{(\mathbf{Z})}\end{aligned} \quad (3.5)$$

Now from Equation (3.2), since $|\mathbf{U}'\mathbf{S}\mathbf{U}|=|\mathbf{G}|$, we must have

$$\begin{aligned}|\mathbf{S}||\mathbf{U}\mathbf{U}'|&=|\mathbf{G}|\\|\mathbf{U}\mathbf{U}'|&=|\mathbf{G}|\times\frac{1}{|\mathbf{S}|}\end{aligned}$$

Using this result, the left-hand side of Equation (3.5) simplifies as

$$\begin{aligned}\left|\mathbf{U}'\mathbf{S}_{(I_k)}\mathbf{U}\right|&=\left|\mathbf{S}_{(I_k)}\right|\left|\mathbf{U}\mathbf{U}'\right|\\&=|\mathbf{G}|\times\frac{\left|\mathbf{S}_{(I_k)}\right|}{|\mathbf{S}|}\\&=|\mathbf{G}|r_k\end{aligned} \quad (3.6)$$

Therefore, from Equations (3.5) and (3.6), we obtain $r_k^{(\mathbf{Z})}=r_k$.

This result means that irrespective of the dimensionality

of the data matrix X, the statistic, $r_k^{(\mathbf{Z})}$ for testing the k-tuple of observations in the set I_k for discordancy in the transformed data Z is numerically the same as that obtained from the X. In addition, it can be verified that the eigenvectors corresponding to the k non-zero eigenvalues of

$$\mathbf{E}_k = \mathbf{S}^{-1}\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)$$

represent suitable dimensions along which the k-tuple of outliers can be optimally highlighted.

3.1. Remarks on Effect of Dimensionality on Multiple Outlier Display

In the derivation in this section, it has been noted that the product matrix $\mathbf{U}'\mathbf{S}\mathbf{U}$ is generally of the form

$$\mathbf{U}'\mathbf{S}\mathbf{U} = \left(\begin{array}{c|c} \mathbf{I}_k & \mathbf{0} \\ \hline \mathbf{0}' & \mathbf{G} \end{array}\right)$$

It is clear that if the dimensionality of the dataset is equal to $(k+1)$ then the matrix is diagonal and $\mathbf{G} = 1$. In this case, $\mathbf{U}'\mathbf{S}\mathbf{U} = \mathbf{I}_{k+1}$ which simplifies the proof. It is this case that has been shown [4] for $k = 2$ using a 3-dimensional dataset. The proof also establishes a basic condition that the number of outliers that can be highlighted on displaying components cannot exceed the dimensionality of the dataset.

4. Illustration

Single Outlier: Suppose $\mathbf{x}_\varepsilon$ is the single outlier. Then $\mathbf{E}_1 = \mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})'$. It can be shown that the eigenvector corresponding to the single non-zero eigenvalue of $\mathbf{E}_1$ is $\mathbf{V}_1 = \mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})$. which is the 1-ODC. Consider the projection of the mean-corrected data onto $\mathbf{V}_1$. This yields the univariate data $(\mathbf{X} - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})$. Now,

$$\max_{i=1,\,2,\,\cdots,\,n}(\mathbf{x}_i - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}}) = (\mathbf{x}_\varepsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})$$

The variation in the univariate projection $\mathbf{Y} = \mathbf{X} * \mathbf{V}_1$ is

$$\begin{aligned} Var(\mathbf{Y}) &= \mathbf{V}_1'\mathbf{S}\mathbf{V} \\ &= (\mathbf{x}_\varepsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}\mathbf{S}\left[\mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})\right] \\ &= (\mathbf{x}_\varepsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}}) \end{aligned}$$

which is the same as the generalised distance of the outlier from the sample mean.

Therefore the Wilks' one-outlier statistic given by

$$\frac{|\mathbf{S}_{(I_k)}|}{|\mathbf{S}|} = 1 - \frac{n}{n-1}(\mathbf{x}_n - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_n - \bar{\mathbf{x}})$$

is therefore equivalent to using the statistic

$$D_{(\varepsilon)} = (n-1)(\mathbf{x}_\varepsilon - \bar{\mathbf{x}})'\mathbf{S}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}}).$$

which is compare to values of Table XXXII by Barnett and Lewis [2], assuming multivariate normality. Thus, the outlier is that observation whose removal optimally reduces the variation in the univariate data. This is a lot simplification of repeated use of the Wilks' ratio.

The plots in Fig. 1 is the projection of the U.S. Food Price data [11] on 1-ODC and a modified 1-ODC showing observation 10 as the single outlier.

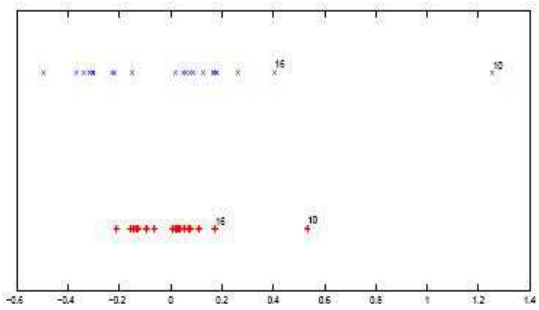

Figure 1: Projection of Food Price data on Original (in '*') and Modified (in '×') 1-ODCs

It has been shown [8] that the actual separation of the single outlier is more effectively highlighted along a modified component $\beta_\varepsilon = \mathbf{S}_{(\varepsilon)}^{-1}(\mathbf{x}_\varepsilon - \bar{\mathbf{x}}_{(\varepsilon)})$. (shown in '×' in Fig 1) which rather involves the mean $\bar{\mathbf{x}}_{(\varepsilon)}$ of the remaining observations when $\mathbf{x}_\varepsilon$ is deleted from the dataset and its corresponding SSCP, $\mathbf{S}_{(\varepsilon)}$. In this dataset, it can be verified that observation 10 is a discordant outlier at 5 percent level of significance.

Outlier-pair: The Milk Transportation Cost data ([5]; [1]) is used in the case for $k = 2$. In this dataset, the pair of observations {9, 21} are known to be outliers. Fig. 2 gives the plot of the data along the two eigenvectors (2-ODC) corresponding to the two non-zero eigenvalues of the matrix $\mathbf{E}_2$.

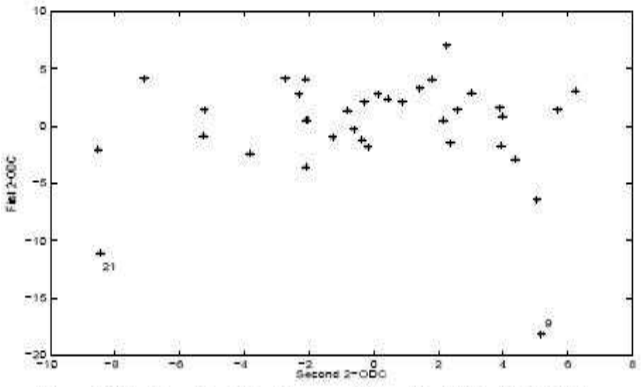

Figure 2: Scatter-plot of the Transportation Cost Data Highlighting Two Outliers in Two Dimensions

Outlier-triple: The Iris Virginica dataset (see e.g. [5]; [1]) is used in this case for $k=3$. In this dataset, observations {19, 18, 32} are known to be the outlier-triple. Fig. 3 gives the projection of the data along the three eigenvectors (3-ODC) corresponding to the three non-zero eigenvalues of the matrix $\mathbf{E}_3$. It must be pointed out that generally, a truly distinct observation in three dimensional space may be difficult to observe. The coordinates of the point in space could actually reveal the real position of the point.

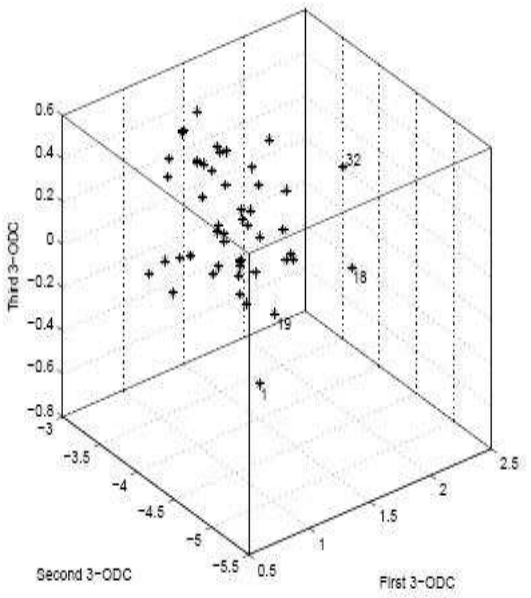

Figure 3: Plot of Projected Iris Virginica Data in Three Dimensions

Although observation 1 in Fig. 3 appears distinct, it is not a member of the outlier triple by our method. It can be verified that in both datasets, the Wilks' multiple-outlier statistic in the reduced dimensions is equal to that obtained in the original dataset.

5. Conclusion

The paper has established two main results. One of the results was derivation of an updating formula for multiple outlier detection and display. The paper shows that an explicit expression for SSCP matrix $\mathbf{A}_{I_k}$, of the set of k outliers that satisfies the matrix equation $\mathbf{S}_{(I_k)} = \mathbf{S} - \mathbf{A}_{I_k}$ is given as

$$\mathbf{A}_{I_k} = \begin{pmatrix} \dfrac{n}{n-k}\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})' \\ -\dfrac{k}{n-k}\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}}_{I_k})' \end{pmatrix}$$

The expression suggests that unlike a single outlier, multiple outliers could be more difficult to detect. In the multiple outlier case, the formula shows that one needs to take into consideration the relative position of each observation from the total sample mean as well as the position from the mean of the set of outliers.

Using the updating formula, it has been shown that in general the discordancy of *k*-tuple outliers is preserved along *k*-Outlier Displaying Components with much lower dimensions than the original high-dimensional dataset. The displaying components are the eigenvectors corresponding to the *k* non-zero eigenvalues of the matrix

$$\mathbf{E}_k = \mathbf{S}^{-1}\left(\sum_{j\in I}(\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'\right)$$

where $I_k = \{i_1, i_2, \cdots, i_k\}$ is an indexed set of labelled most extreme k-tuple of observations in the dataset.

Appendix

Remarks on the Properties of the $\mathbf{C}_2$ Matrix

The matrix $\mathbf{C}_2$ in Equation (2.3) has a number of interesting properties. In the following theorem, we state and prove these properties.

Theorem

The matrix given by

$$\mathbf{C}_2 = \left(\begin{array}{c|c} \left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k \\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k} & \mathbf{I}_k - \frac{1}{n}\mathbf{1}_k\mathbf{1}'_k \end{array}\right)$$

Satisfies the following the following properties:

$$\mathbf{C}_2' = \mathbf{C}_2$$

$$\mathbf{C}_2^2 = \mathbf{C}_2$$

$$\mathbf{C}_1\mathbf{C}_2 = \mathbf{0}_n$$

$$\mathbf{C}_2(\mathbf{1}_n\bar{\mathbf{x}}') = 0$$

Proof

Property 1: Clearly $\mathbf{C}_2' = \mathbf{C}_2$.

Property 2

$$\mathbf{C}_2^2 = \left(\begin{array}{c|c} \left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k \\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k} & \mathbf{I}_k - \frac{1}{n}\mathbf{1}_k\mathbf{1}'_k \end{array}\right)\left(\begin{array}{c|c} \left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k \\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k} & \mathbf{I}_k - \frac{1}{n}\mathbf{1}_k\mathbf{1}'_k \end{array}\right)$$

Let the general element of $\mathbf{C}_2^2$ be $(\mathbf{C}_{ij})$. Then

$$\begin{aligned} \mathbf{C}_{11} &= \left(-\frac{1}{n}+\frac{1}{n-k}\right)^2(\mathbf{1}_{n-k}\mathbf{1}'_{n-k})(\mathbf{1}_{n-k}\mathbf{1}'_{n-k}) + \frac{1}{n^2}(\mathbf{1}_{n-k}\mathbf{1}'_k)(\mathbf{1}_k\mathbf{1}'_{n-k}) \\ &= \frac{k^2}{n^2(n-k)}\mathbf{1}_{n-k}\mathbf{1}'_{n-k} + \frac{k}{n^2}\mathbf{1}_{n-k}\mathbf{1}'_{n-k} \\ &= \frac{k}{n(n-k)}\mathbf{1}_{n-k}\mathbf{1}'_{n-k} \\ &= \left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} \end{aligned}$$

$$\begin{aligned}\mathbf{C}_{12} &= -\frac{1}{n}\left(-\frac{1}{n}+\frac{1}{n-k}\right)(\mathbf{1}_{n-k}\mathbf{1}'_{n-k})(\mathbf{1}_{n-k}\mathbf{1}'_k)-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k+\frac{1}{n^2}(\mathbf{1}_{n-k}\mathbf{1}'_k)(\mathbf{1}_k\mathbf{1}'_k)\\ &= -\frac{k}{n^2(n-k)}(n-k)\mathbf{1}_{n-k}\mathbf{1}'_k+\frac{k}{n^2}\mathbf{1}_{n-k}\mathbf{1}'_k-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\\ &= -\frac{k}{n^2}\mathbf{1}_{n-k}\mathbf{1}'_k+\frac{k}{n^2}\mathbf{1}_{n-k}\mathbf{1}'_k-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\\ &= -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\end{aligned}$$

$\mathbf{C}_{21}$ can be shown similarly to be equal to $\mathbf{C}_{12}$. Now taking $\mathbf{C}_{22}$ we obtain

$$\begin{aligned}\mathbf{C}_{22} &= \frac{1}{n^2}(\mathbf{1}_k\mathbf{1}'_{n-k})(\mathbf{1}_{n-k}\mathbf{1}'_k)+\left(\mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}'_k\right)^2\\ &= \frac{n-k}{n^2}\mathbf{1}_k\mathbf{1}'_k+\mathbf{I}_k-\frac{2}{n}\mathbf{1}_k\mathbf{1}'_k+\frac{1}{n^2}(\mathbf{1}_k\mathbf{1}'_k)(\mathbf{1}_k\mathbf{1}'_k)\\ &= \frac{n-k}{n^2}\mathbf{1}_k\mathbf{1}'_k+\mathbf{I}_k-\frac{2}{n}\mathbf{1}_k\mathbf{1}'_k+\frac{k}{n^2}\mathbf{1}_k\mathbf{1}'_k\\ &= \frac{1}{n}\mathbf{1}_k\mathbf{1}'_k-\frac{k}{n^2}\mathbf{1}_k\mathbf{1}'_k+\mathbf{I}_k-\frac{2}{n}\mathbf{1}_k\mathbf{1}'_k+\frac{k}{n^2}\mathbf{1}_k\mathbf{1}'_k\\ &= \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}'_k\end{aligned}$$

Thus,

$$\mathbf{C}_2^2=\left(\begin{array}{c|c}\left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k} & \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}'_k\end{array}\right)$$

which is the same as $\mathbf{C}_2$. Therefore, $\mathbf{C}_2^2=\mathbf{C}_2$.

Property 3

$$\mathbf{C}_1\mathbf{C}_2=\left(\begin{array}{c|c}\mathbf{I}_{n-k}-\frac{1}{n-k}\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & \mathbf{0}_{(n-k)\times k}\\ \hline \mathbf{0}_{k\times(n-k)} & \mathbf{0}_{k\times k}\end{array}\right)\left(\begin{array}{c|c}\left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k} & \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}'_k\end{array}\right)$$

Let the general element of this product be $(\mathbf{t}_{ij})$.

$$\begin{aligned}\mathbf{t}_{11} &= \frac{k}{n(n-k)}\left(\mathbf{I}_{n-k}-\frac{k}{n-k}\mathbf{1}_{n-k}\mathbf{1}'_{n-k}\right)(\mathbf{1}_{n-k}\mathbf{1}'_{n-k})\\ &= \frac{k}{n(n-k)}\left(\mathbf{1}_{n-k}\mathbf{1}'_{n-k}-\frac{1}{n-k}(\mathbf{1}_{n-k}\mathbf{1}'_{n-k})(\mathbf{1}_{n-k}\mathbf{1}'_{n-k})\right)\\ &= \frac{k}{n(n-k)}(\mathbf{1}_{n-k}\mathbf{1}'_{n-k}-\mathbf{1}_{n-k}\mathbf{1}'_{n-k})\\ &= \mathbf{0}_n\end{aligned}$$

$$\begin{aligned}\mathbf{t}_{12} &= -\frac{1}{n}\left(\mathbf{I}_{n-k}-\frac{1}{n-k}\mathbf{1}_{n-k}\mathbf{1}'_{n-k}\right)(\mathbf{1}_{n-k}\mathbf{1}'_k)\\ &= -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k+\frac{1}{n(n-k)}(\mathbf{1}_{n-k}\mathbf{1}'_{n-k})(\mathbf{1}_{n-k}\mathbf{1}'_k)\\ &= -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k+\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\\ &= \mathbf{0}_n\end{aligned}$$

It can be shown similarly that $\mathbf{t}_{21}=\mathbf{t}_{22}=\mathbf{0}_n$.

Therefore, $\mathbf{C}_1\mathbf{C}_2=\mathbf{0}_{n\times n}$.

Property 4

Let $\bar{\mathbf{x}}=(\bar{\mathbf{x}}_1,\bar{\mathbf{x}}_2,\cdots,\bar{\mathbf{x}}_p)$ be the p-dimensional mean vector. Then consider the $n\times p$ matrix

$$\mathbf{1}_n\bar{\mathbf{x}}'=\left(\bar{\mathbf{x}}_1\mathbf{1}_n\mid\bar{\mathbf{x}}_2\mathbf{1}_n\mid\cdots\mid\bar{\mathbf{x}}_i\mathbf{1}_n\mid\cdots\mid\bar{\mathbf{x}}_p\mathbf{1}_n\right)$$

Now, partition each $\mathbf{1}_n\in\Re^n$ in the form $\mathbf{1}_n=(\mathbf{1}_{n-k}\mid\mathbf{1}_k)'$. Thus,

$$\mathbf{1}_n\bar{\mathbf{x}}=\left(\begin{array}{c|c|c|c|c|c}\bar{\mathbf{x}}_1\mathbf{1}_{n-k} & \bar{\mathbf{x}}_2\mathbf{1}_{n-k} & \cdots & \bar{\mathbf{x}}_i\mathbf{1}_{n-k} & \cdots & \bar{\mathbf{x}}_p\mathbf{1}_{n-k}\\ \hline \bar{\mathbf{x}}_1\mathbf{1}_k & \bar{\mathbf{x}}_2\mathbf{1}_k & \cdots & \bar{\mathbf{x}}_i\mathbf{1}_k & \cdots & \bar{\mathbf{x}}_p\mathbf{1}_k\end{array}\right)$$

Thus,

$$\mathbf{C}_2(\mathbf{1}_n\bar{\mathbf{x}}')=\left(\begin{array}{c|c}\left(-\frac{1}{n}+\frac{1}{n-k}\right)\mathbf{1}_{n-k}\mathbf{1}'_{n-k} & -\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k} & \mathbf{I}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}'_k\end{array}\right)\times\left(\begin{array}{c|c|c|c|c|c}\bar{\mathbf{x}}_1\mathbf{1}_{n-k} & \bar{\mathbf{x}}_2\mathbf{1}_{n-k} & \cdots & \bar{\mathbf{x}}_i\mathbf{1}_{n-k} & \cdots & \bar{\mathbf{x}}_p\mathbf{1}_{n-k}\\ \hline \bar{\mathbf{x}}_1\mathbf{1}_k & \bar{\mathbf{x}}_2\mathbf{1}_k & \cdots & \bar{\mathbf{x}}_i\mathbf{1}_k & \cdots & \bar{\mathbf{x}}_p\mathbf{1}_k\end{array}\right)$$

The product with the ith column vector of $\mathbf{1}_n\bar{\mathbf{x}}'$ gives

$$\begin{aligned}&\mathbf{x}_i\left(\begin{array}{c}\frac{k}{n(n-k)}\mathbf{1}_{n-k}\mathbf{1}'_{n-k}\mathbf{1}_{n-k}-\frac{1}{n}\mathbf{1}_{n-k}\mathbf{1}'_k\mathbf{1}_k\\ \hline -\frac{1}{n}\mathbf{1}_k\mathbf{1}'_{n-k}\mathbf{1}_{n-k}+\mathbf{1}_k-\frac{1}{n}\mathbf{1}_k\mathbf{1}'_k\mathbf{1}_k\end{array}\right)\\ &=\mathbf{x}_i\left(\begin{array}{c}\frac{k}{n}\mathbf{1}_{n-k}-\frac{k}{n}\mathbf{1}_{n-k}\\ \hline -\frac{n-k}{n}\mathbf{1}_k+\mathbf{1}_k-\frac{k}{n}\mathbf{1}_k\end{array}\right)=\mathbf{x}_i\left(\begin{array}{c}\mathbf{0}_{n-k}\\ \hline \mathbf{0}_k\end{array}\right)\end{aligned}$$

Therefore, $\mathbf{C}_2(\mathbf{1}_n\bar{\mathbf{x}}')=0$.

References

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

[2] Barnett, V., & Lewis, T. (1994). Outliers in Statistical Data. (3rd ed.) New York: John Wiley and Sons Limited.

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

[4] Gordor, B. K. & Fieller, N. R. J. (1999). How to display an outlier in multivariate datasets. Journal of Applied Sciences & Technology, 4(2).

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

[6] Miller, K. S. (1981). On the Inverse of the Sum of Matrices. JSTOR 54(2), 67-72.

[7] Nkansah, B. K. & Gordor B. K. (2012a): A Procedure for Detecting a Pair of Outliers in Multivariate Datasets. Studies in Mathematical Sciences 4(2), 1-9.

[8] Nkansah, B. K. & Gordor B. K. (2012b): On the One-Outlier Displaying Component in Multivariate Datasets. Journal of Informatics and Mathematical Sciences, 4(2), 229-239.

[9] Pan, J. X., & Wang, X. R. (1994). Unbiasedness of a Multivariate Outlier Test For Elliptically Contoured Distributions, Multivariate Analysis and its Applications. Monograph Series, 24, 457-460.

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

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

[12] Wilks, S. S. (1963). Multivariate Statistical Outliers. Sankhya, A, 25, 407-426.