

Full Length Research Paper

A robust method of estimating covariance matrix in multivariate data analysis

G. M. Oyeyemi* and R. A. Ipinyomi

Department of Statistics, University of Ilorin, Nigeria.

Accepted 25 September, 2009

We proposed a robust method of estimating covariance matrix in multivariate data set. The goal is to compare the proposed method with the most widely used robust methods (Minimum Volume Ellipsoid and Minimum Covariance Determinant) and the classical method (MLE) in detection of outliers at different levels and magnitude of outliers. The proposed robust method competes favourably well with both MVE and MCD and performed better than any of the two methods in detection of single or fewer outliers especially for small sample size and when the magnitude of outliers is relatively small.

Key words: Covariance matrix, minimum volume ellipsoid (MVE), minimum covariance determinant (MCD), mahalanobis distance, optimality criteria.

INTRODUCTION

Let $X = \{x_1, x_2, x_3, \dots, x_m\}$ be a set of m points in \mathfrak{R}^p , where x_i , $i = 1, 2, 3, \dots, m$, are independent and identically distributed multivariate normal $N_p(\mu, \Sigma)$. The usual Maximum Likelihood Estimate (MLE) method of estimating μ and Σ are the sample mean vector and sample covariance matrix \bar{x} and S , respectively which is given as;

$$\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i = [\bar{x}_1, \bar{x}_2, \dots, \bar{x}_p]$$

$$S = \frac{1}{m-1} \sum_{i=1}^m (x_i - \bar{x})(x_i - \bar{x})', \text{ that is}$$

$$S = \begin{bmatrix} s_{11} & s_{12} & s_{1j} & \dots & s_{1p} \\ s_{21} & s_{22} & s_{2j} & \dots & s_{2p} \\ \dots & \dots & s_{ij} & \dots & \dots \\ \dots & \dots & \dots & \dots & \dots \\ s_{p1} & s_{p2} & s_{pj} & \dots & s_{pp} \end{bmatrix}$$

Covariance matrix plays a prominent role in multivariate data analysis. It measures the spread of the individual variables as well as the level of inter-relationship (inter-corelation) that may exist between pairs of the variables in the multivariate data. The statistic is used in many multivariate techniques as a measure of spread and inter-corelation among variables. Such techniques include multivariate regression model, Principal Component and Factor Analyses, Canonical Correlation and Linear Discriminant and even in Multivariate Statistical Process Control (MSPC). It is used in obtaining multivariate control charts such as Hotelling T^2 chart, Multivariate Exponential Moving Weighted Average (MEMWA) chart and Multivariate Cumulative Sum (MCUSUM) chart.

It is well known that the Maximum Likelihood Estimate (MLE) method can be very sensitive to deviations from the assumptions made on the data, in particular, to unexpected outliers in the data (Vandev and Neykov, 2000). To overcome this problem, many robust alternatives to Maximum Likelihood Estimator (MLE) have been developed in recent years. All the methods converged on tackling the problem of robust estimation by finding a sufficiently large subset of uncontaminated (free of outliers) datas. Such subset will be mainly elements of the true population and estimation is then based on this subset.

When the estimate of the covariance matrix of a multivariate data is not robust (Biased), it often times renders the technique or the analysis it is used for to be inconsistent and in most cases found to be unreliable.

*Corresponding author. E-mail: gmoyeyemi@yahoo.com, ipinyomira@yahoo.co.uk.

For instance, in Multivariate Control chart like Hotelling T^2 chart, where the presence of multiple outliers can affect the estimation of the covariance matrix, using Maximum Likelihood Estimation (MLE) method to the extent that all the outliers will go undetected by the chart (Vargas 2003). The same effect can be experienced in other analyses like Principal Component and Factor analysis and other multivariate techniques.

Outliers can heavily influence the estimation of the covariance matrix Σ and subsequently the parameters or statistics that are needed to be derived from it. Hence, a robust estimate of the covariance matrix that will not be affected by outliers is required to obtain valid and reliable results (Hubert and Engelen, 2007). There have been many robust methods of estimating the covariance matrix of a multivariate data. Such methods include Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD), S-Estimator, M-Estimator and Orthogonalized Gnanadesikan-Kettering (OGK) methods.

This paper gives a brief overview of the most widely used methods (Minimum Volume Ellipsoid and Minimum Covariance Determinant) and also introduces our robust estimation method. We compared our robust method with the two methods based on both simulated and real life multivariate data in detection of outliers as bases for comparison.

ROBUST METHODS

Minimum volume ellipsoid (MVE)

The Minimum Volume Ellipsoid (MVE) estimator was first proposed by Rousseeuw (1984). It has been studied extensively for non-control chart settings and frequently used in detection of multivariate outliers. The estimation seeks to find the ellipsoid of minimum volume that covers a subset of at least h data points. The subset of size h is called halfset because h is often chosen to be just more than half of the m data points. The location estimator is the geometrical center of the ellipsoid and the estimator of the variance-covariance matrix defining the ellipsoid itself multiplied by an appropriate constant to ensure consistency (Rousseeuw and van Zomeren, 1990; Rousseeuw and van Zomeren, 1991; Rocke and Woodruff, 1998).

Assuming that we have a multivariate data set containing m samples, $\{x_i \in \mathfrak{R}^p\}_{i=1}^m$. In order to solve the MVE problem, we need to obtain a $p \times p$ positive definite matrix $C \in \mathfrak{R}^{p \times p}$ and the center of the ellipsoid t so as to maximize $\det(C^{-1})$ subject to $(x_i - t)^T C^{-1} (x_i - t) \leq p$ (Titterington, 1975).

The MVE for the data set $\{x_i\}_{i=1}^m$ must go through at least $p+1$ and at most h support vectors. Thus, the MVE estimates of the location and dispersion do not correspond to the sample mean vector and sample variance-

covariance matrix of a particular halfset. For more detailed discussion on MVE see Daves (1987), Lopuhaa and Rousseeuw (1991), Titterington (1975) and (Agullo, 1996).

Minimum covariance determinant (MCD)

An alternative high breakdown estimation procedure to the MVE is an estimator based on the Minimum Covariance Determinant (MCD), which was first proposed by Rousseeuw (1984). It is obtained by finding the halfset of multivariate data points that gives the minimum value of the determinant of the covariance matrix. The resulting estimator of location is the sample mean vector of the points that is the halfset and the estimator of the dispersion is the sample covariance matrix of the points multiplied by an appropriate constant to ensure consistency just as was done for MVE.

The MCD estimators are intuitively appealing because a small value of the determinant corresponds to near linear dependencies of the data in the p -dimensional space that is because a small determinant corresponds to a small Eigenvalue which suggests a near linear dependency that suggests that there is a group of points that are similar to each other (Jensen et al., 2002).

Let $p < m/2$, let $X = \{x_1, x_2, x_3, \dots, x_p\}$ be a set of

n points in \mathfrak{R}^p . Let h be a natural number, $m/2 < h < m$. The Minimum Covariance Determinant problem for X and h , MCD for short, is the problem to find an h -elements set $X^h = \{x_{i_1}, x_{i_2}, \dots, x_{i_h}\} \in X$ such that $\det(X^h)$ is the minimal overall h -element sets. The empirical covariance matrix $C(X^h)$, with minimal determinant yields a robust estimate S of the scatter matrix, with $S = S(X^h) = C_0 C(X^h)$, where C_0 is a suitably chosen constant to achieve consistency. The estimate of the location parameter is given as;

$$t = t(X^h) = \frac{1}{h} \sum_{x \in X^h} x$$

The pair (t, S) is called the MCD-estimate with respect to X .

THE PROPOSED ROBUST METHOD (PRM)

Given a p -dimensional multivariate normal data $X_{p \times m}$ with m observations $\{x_i\}_{i=1}^m$. Our interest is to obtain a subset of $\{x_i\}_{i=1}^m$ of size $k = p+1$ that will satisfy some optimality criteria. Therefore, we sample without replacement a sample of size k from m , this will give C_{p+1}^m possible subsets of size $p+1$, $x_{j_1}, x_{j_2}, \dots, x_{j_{p+1}}, \{x_j\}_{j=1}^{p+1}$. If each

subset is denoted by $J_j, j = 1, 2, \dots, C_{p+1}^m$. For each J_j , we estimate the variance-covariance matrix, C_j ;

$$C_j = \frac{1}{p+1} (x_j - \bar{x}_j)(x_j - \bar{x}_j)^T$$

And for each of the $p \times p$ matrix C_j , the characteristic roots or Eigenvalues $e_{j1}, e_{j2}, \dots, e_{jp}$ are obtained and from such Eigenvalues, the following optimality criteria can be calculated;

E_A = The Minimum of the minimum Eigenvalues.
 $E_A = \min\{\min e_i\}$

E_P = The Minimum of the product of the Eigenvalues.
 $E_P = \min\left\{\prod_{i=1}^p e_i\right\}$

E_H = The Minimum of the harmonic mean of the Eigenvalues. $E_H = \min\left\{\sum_{i=1}^p \frac{1}{e_i}\right\}^{-1}$

The objective is to obtain data points $(p+1)$ such that its variance-covariance matrix will satisfy all the three optimality criteria. Such covariance matrix will be inflated or deflated to accommodate good data points among the observed data. The resulting variance-covariance matrix is then multiplied by a constant for consistency.

THE ALGORITHM FOR OBTAINING THE PROPOSED METHOD

Let $X = \{x_1, x_2, x_3, \dots, x_m\}$ be a set of m points in \mathcal{R}^p . Let h be a natural number such that $m/2 < h < m$.

The $p+1$ data points $\{x_1, x_2, \dots, x_{p+1}\}$, that satisfy the three optimality criteria were selected and use to obtain the center and variance-covariance matrix;

$$\bar{x}_* = \frac{1}{p+1} \sum_{i=1}^p x_{i*} \text{ and } S_* = \frac{1}{p+1} (x_* - \bar{x}_*)(x_* - \bar{x}_*)^T, \text{ respectively.}$$

The mahalanobis distances for all the m observations using \bar{x}_* and S_* can be calculated as follows;

$$d_i^2 = (x - \bar{x}_*)S_*^{-1}(x - \bar{x}_*)^T$$

The d_i^2 ($i = 1, 2, 3, \dots, m$) is arranged in order of magnitude from the least to the highest. The first $p+j$ ($j =$

2, 3, 4, , , , $h-p-1$) distances are selected and their corresponding sample units (points) are used to compute the next \bar{x}_* and S_* as follows; $\bar{x}_* = \frac{1}{p+j} \sum_{i=1}^p x_{i*}$ and

$$S_* = \frac{1}{p+j} (x_* - \bar{x}_*)(x_* - \bar{x}_*)^T, \text{ respectively.}$$

The new set of \bar{x}_* and S_* are then used to obtain the mahalanobis distances for all the observations.

Steps 4, 5 and 6 are repeated until the number of units selected is $h = \frac{m+p+1}{2}$. The Proposed robust estimators are then given as;

$$\bar{x}_{Prop} = \frac{1}{h} \sum_{i=1}^h x_{i*} \text{ and } S_{Prop} = \frac{1}{h} (x_* - \bar{x}_*)(x_* - \bar{x}_*)^T$$

Where: $x_* = \{x_1, x_2, \dots, x_h\}$.

Simulation

For the purpose of comparing the proposed robust method with other methods, Monte Carlo simulation was adopted to generate the sets of bivariate normal samples. Also, the same procedure was used to obtain the upper control limits for all the four methods under comparison. A set of $m = 30$ observations was generated from a bivariate normal distribution. The proposed robust method was compared with the other three methods (Minimum Volume Ellipsoid, Minimum Covariance Determinant and The Classical). We assume the Non-centrality Parameter $ncp = (\mu - \mu_0)' \Sigma^{-1} (\mu - \mu_0)$ to be the measure of severity of a shift to the out-of-control mean vector $\underline{\mu}$ from the in-control mean vector $\underline{\mu}_0$. And because the signal probability depends on the value of the non-centrality parameter but not on the in-control mean vector $\underline{\mu}_0$ or the variance-covariance matrix Σ , we made use, without loss of generality, the zero vector as $\underline{\mu}_0$ and the identity matrix of order two, I_2 , as Σ .

The control limits were determined from 5000 simulations, such that all the methods considered had overall false alarm probability of 0.05. The limits were obtained by generating 5000 data set for m and p ($m = 30$ and $p = 2$). The Hotelling- T^2 statistic, T_i^2 , were computed for $i = 1, 2, 3, \dots, m$. The maximum value was recorded and the 95th percentile of the maximum values of the Hotelling's - T^2 for $j = 1, 2, 3, \dots, 5000$ was taken to be the Upper Control Limit (UCL) for the control chart. The values obtained were 9.686, 38.166, 33.917 and 63.326 for the Classical, MVE, MCD and Proposed methods, respectively.

Table 1. The signal probability when there is one outlier.

NCP	Methods			
	Classical	MVE	MVD	Proposed
5	0.1100	0.0450	0.0550	0.0750
10	0.4100	0.1300	0.1400	0.2500
15	0.5500	0.3000	0.3100	0.4200
20	0.7200	0.4200	0.4000	0.5100
25	0.8100	0.5300	0.5500	0.6750
30	0.9400	0.7600	0.7300	0.8250

Table 2. The signal probability when there are 3 outliers.

NCP	Methods			
	Classical	MVE	MVD	Proposed
5	0.0430	0.0750	0.0285	0.0400
10	0.1000	0.0870	0.0930	0.1330
15	0.1200	0.2130	0.2070	0.2470
20	0.1600	0.3350	0.3450	0.3550
25	0.1600	0.4670	0.4450	0.4300
30	0.1900	0.5830	0.5770	0.5270

Table 3. The signal probability when there are 5 outliers.

NCP	Methods			
	Classical	MVE	MVD	Proposed
5	0.0100	0.0080	0.0080	0.0240
10	0.0110	0.0420	0.036	0.0560
15	0.0100	0.1180	0.0980	0.1120
20	0.0200	0.2280	0.2140	0.2240
25	0.0180	0.3580	0.3500	0.2940
30	0.0180	0.3660	0.3700	0.3080

The Lower Control Limit is always set to zero.

Once the control limits are set, k ($k = 1, 3, 5$ and 7) outliers are randomly generated among the m ($m = 30$) observations. To generate the outliers, the process mean vector was changed from $\mu = \mu_0$ to $\mu = \mu_1$ to obtain a given value of non-centrality parameter. The charts were compared by estimating the probability of obtaining a valid signal. These probabilities were calculated from 1000 replications. The illustrations were made for $k = 1, 3, 5$ and 7 as shown in Tables 4 to 7.

Figures 1 to 4 show the estimated signal probabilities for different non-centrality parameter values ($npc = 5, 10, 15, 20, 25$ and 30). When there is only one outlier, it can be seen that the control chart based on Classical method is effective in detecting the outlier than the other three

Table 4. The signal probability when there are 7 outliers.

NCP	Methods			
	Classical	MVE	MVD	Proposed
5	0.0100	0.0165	0.0170	0.0210
10	0.0120	0.0450	0.4800	0.0470
15	0.0070	0.0900	0.0840	0.0840
20	0.0100	0.1560	0.147	0.1300
25	0.0100	0.1980	0.2000	0.1820
30	0.0030	0.2610	0.2340	0.1960

methods. Figure 1 shows that the line of Classical method reaches a probability value equal or greater than 0.9 when the non-centrality parameter is 30. Though the other methods are less powerful for a single outlier but they still signal that there exist outliers with a reasonable probability with the proposed method having an edge over the other two methods (MVE and MCD).

However, for multiple outliers, the Classical Control Chart performed poorly in detecting outliers. The method becomes worst when there are 5 or 7 outliers in the data set. For instance, when there were 3 outliers and the non-centrality parameter is 20, the estimated signal probability was only 0.1600 for the Classical Control Chart while the signal probabilities for MVE, MCD and Proposed Control Charts were 0.423, 0.413 and 0.417, respectively. For $k = 5$ or 7 outliers, the margin between the Classical and the other control charts became more pronounced. For instance, when $k = 7$ and non-centrality parameter value is 30, the signal probability value for Classical Chart is 0.003 while MVE, MCD and Proposed Control Chart have signal probability values of 0.261, 0.234 and 0.244, respectively.

Generally, from the control charts, it can be inferred that for all the methods, as the number of outliers increases so the signal probability decreases for a given sample size. Also, for all the control charts except the Classical chart, as the value of non-centrality parameter increases, the signal probability values increase. As a result of this, it can be inferred that the Classical Control Chart is optimized for none or single outlier while the other robust methods were optimized for detecting multiple outliers, usually the number of outliers should be less than $m-p-1/2$ where p is the number of variables and m is the number of observations (Vargas, 2003).

Table 1 gives the signal probability of the four methods for a single outlier for varying size of non-centrality parameter. For NCP value of 5, the classical chart has the highest probability value of 0.1100 closely followed by the proposed chart with probability of 0.075 while MCD and MVE have probability values of 0.0550 and 0.0450, respectively. When the NCP is 30, the signal probability values for all the four charts are 0.9400, 0.8250, 0.7600 and 0.7300 for classical, proposed, MVE and MCD

Table 5. Data set and Hotelling's-T2 statistic using the Classical, MCD, MVE and the Proposed Robust Method (PRM) when there is only one (1) outlier.

No.	X ₁	X ₂	Classical	MCD	MVE	Proposed
1	0.567	60.558	0.8066	0.8463	1.1238	0.8038
2	0.538	56.303	12.9755	27.0895	67.2475	102.0470
3	0.530	59.524	0.1373	0.6041	1.0070	3.4001
4	0.562	61.102	1.8375	2.5720	4.9464	2.2175
5	0.483	59.834	1.5697	2.2028	1.5167	0.9069
6	0.525	60.228	0.3301	0.3615	0.5323	0.0033
7	0.556	60.756	0.9772	1.2021	2.2802	0.7247
8	0.586	59.823	0.9045	1.8130	1.6442	8.0729
9	0.547	60.153	0.1269	0.0509	0.0689	0.6943
10	0.531	60.640	0.8008	0.9546	2.2730	0.8574
11	0.581	59.785	0.7192	0.6552	1.5253	7.6383
12	0.585	59.675	0.9097	0.9091	2.2835	9.8820
13	0.540	60.489	0.4835	0.5115	1.1256	0.1808
14	0.458	61.067	5.2413	6.2687	13.3162	17.1425
15	0.554	59.788	0.0736	0.1033	0.4754	3.5894
16	0.469	58.640	3.5357	6.9488	7.4692	7.7120
17	0.471	59.574	2.2696	3.4022	2.1857	1.4820
18	0.457	59.718	3.2442	4.3967	3.1098	2.5300
19	0.565	60.901	1.3981	1.8172	3.1419	1.1747
20	0.664	60.180	6.8326	7.0948	7.2069	22.1173
21	0.600	60.493	1.8978	1.9560	1.6148	4.0007
22	0.586	58.370	3.3564	5.8965	17.8616	39.6793
23	0.567	60.216	0.4275	0.3020	0.2946	1.8933
24	0.496	60.214	1.1838	1.4432	1.7373	1.0097
25	0.485	59.500	1.4968	2.5331	1.6721	1.3624
26	0.573	60.052	0.4843	0.3454	0.4769	3.6558
27	0.520	59.501	0.2899	0.8866	1.0456	2.7191
28	0.556	58.476	2.0635	4.5803	12.8536	26.7262
29	0.539	58.666	1.3860	3.5429	8.8320	17.8671
30	0.554	60.239	0.2404	0.1447	0.1892	0.8019

The bold numbers indicate outlying points.

charts, respectively.

Real life data illustration

We consider the data presented in Quesenberry (2001). The data consists of 11 quality characteristics (variables) measured on 30 products from a production process. The first two variables are considered and they are reproduced in columns 2 and 3 of Table 1. The two variables are used to compare the four methods of constructing Hotelling's-T² Control Chart. The sample mean vector and covariance matrix for the unmodified data (Classical method) of the Table 1 are;

$$\bar{X}_{Classical} = \begin{bmatrix} 0.5415 \\ 59.8155 \end{bmatrix}$$

$$S_{Classical} = \begin{bmatrix} 0.002203 & 0.000399 \\ 0.000399 & 0.955066 \end{bmatrix}$$

The location and covariance matrices for the two robust methods (MVE and MCD) using R- Language are given as follows;

Location and Scatter matrix for MVE;

$$\bar{X}_{MVE} = \begin{bmatrix} 0.5419 \\ 60.0200 \end{bmatrix}; S_{MVE} = \begin{bmatrix} 0.002360 & 0.011328 \\ 0.011328 & 0.257772 \end{bmatrix}$$

Location and Scatter matrix for MCD;

$$\bar{X}_{MCD} = \begin{bmatrix} 0.5474 \\ 59.9927 \end{bmatrix};$$

Table 6. Data set and Hotelling's-T² statistic using the Classical, MCD, MVE and the Proposed Robust Method (PRM) when there are three (3) outliers.

No.	X ₁	X ₂	Classical	MCD	MVE	Proposed
1	0.567	60.558	0.2156	0.8208	0.9639	0.6835
2	0.538	56.303	11.1453	26.9371	25.7786	67.2619
3	0.530	59.524	0.1546	0.5949	0.5530	1.6341
4	0.562	61.102	1.1349	2.7306	2.7151	2.0714
5	0.483	59.834	1.4296	3.0245	3.7622	2.0271
6	0.525	60.228	0.6717	0.7466	0.7364	0.1661
7	0.556	60.756	0.6643	1.3542	1.3345	0.7574
8	0.586	59.823	0.2773	0.7474	1.3182	5.7024
9	0.547	60.153	0.1618	0.1293	0.1029	0.2316
10	0.531	60.640	1.1820	1.4730	1.3954	1.0863
11	0.581	59.785	0.2384	0.5831	1.0330	5.2204
12	0.585	59.675	0.4266	0.8712	1.4140	6.8552
13	0.540	60.489	0.6525	0.8012	0.7295	0.2732
14	0.880	65.230	9.0509	107.1500	135.1170	83.6520
15	0.554	59.788	0.0416	0.0503	0.0760	1.9257
16	0.469	58.640	0.9293	6.8119	7.8500	5.5945
17	0.471	59.574	1.4702	4.1925	5.2996	2.7493
18	0.457	59.718	2.3397	5.6685	7.3243	4.7841
19	0.565	60.901	0.6842	1.8775	1.9409	1.1879
20	0.664	60.180	2.1596	7.5567	11.6930	20.7586
21	0.600	60.493	0.0709	1.7452	2.6869	3.6661
22	0.586	58.370	3.7556	6.3072	6.6884	27.0517
23	0.567	60.216	0.0181	0.2386	0.4118	1.1943
24	0.980	66.080	14.6257	165.3657	212.9290	133.3725
25	0.485	59.500	0.9127	2.9998	3.6643	1.6948
26	0.573	60.052	0.0129	0.2515	0.5337	2.4067
27	0.520	59.501	0.2341	0.9372	0.9592	1.3111
28	0.556	58.476	2.0733	4.4845	4.3582	17.0178
29	0.539	58.666	1.1358	3.3007	3.1253	10.7904
30	0.554	60.239	0.1306	0.1814	0.1976	0.3682

The bold numbers indicate outlying points.

$$S_{MCD} = \begin{bmatrix} 0.001943 & -0.000705 \\ -0.000705 & 0.504607 \end{bmatrix}$$

The mean vector and covariance matrix of the data using the proposed robust method (PRM) are as follows;

$$\bar{X}_{Proposed} = \begin{bmatrix} 0.5266 \\ 60.2264 \end{bmatrix}$$

$$S_{Proposed} = \begin{bmatrix} 0.002146 & 0.022083 \\ 0.022083 & 0.387353 \end{bmatrix}$$

The values of Hotelling's-T² statistics, $T_{i,usual}^2$, $T_{i,MCD}^2$, $T_{i,MVE}^2$ and $T_{i,Proposed}^2$ based on the Classical, MCD, MVE

and Proposed robust methods, respectively are presented in columns 4, 5, 6 and 7 of Table 5 in that order.

Comparing the values obtained from the four statistics against their respective upper control limits, which are 9.686, 33.917, 38.166 and 63.326 for the Classical, MCD, MVE and Proposed methods, respectively, it was found that only MCD control chart did not signal the second observation as outlier, the other three methods signaled the second observation as an outlier. Figures 1a through 1d showed the multivariate control charts for the four methods; Classical, MVE, MCD and PRM charts respectively. In MCD - control chart, none of the 30 observations is above the upper control limit (Figure 1c). The other control charts indicated the second observation to be an outlier (out-of-control point).

We arbitrarily introduced two more outlying observations into the data, the observations 14 and 24 were mo-

Table 7. Data set and Hotelling's-T2 statistic using the Classical, MCD, MVE and the Proposed Robust Method (PRM) when there five (5) outliers

No.	X ₁	X ₂	Classical	MCD	MVE	Proposed
1	0.567	60.558	0.2362	0.7804	0.9022	0.6835
2	0.538	56.303	2.9413	33.0764	31.6581	67.2629
3	0.530	59.524	0.1635	0.7412	0.6812	1.6341
4	0.562	61.102	0.7418	2.8987	2.8360	2.0714
5	0.483	59.834	1.4570	2.8809	3.5639	2.0271
6	0.525	60.228	0.6809	0.6788	0.6677	0.1661
7	0.556	60.756	0.5279	1.3746	1.3344	0.7574
8	0.586	59.823	0.0634	0.8143	1.3980	5.7024
9	0.547	60.153	0.2389	0.0890	0.0631	0.2316
10	0.531	60.640	0.9476	1.4706	1.3952	1.0863
11	0.581	59.785	0.0391	0.6687	1.1331	5.2204
12	0.585	59.675	0.0902	1.0170	1.5810	6.8552
13	0.540	60.489	0.5938	0.7650	0.6919	0.2732
14	0.880	65.230	7.4386	110.1994	134.3937	83.6520
15	0.554	59.788	0.0371	0.1161	0.1474	1.9257
16	0.469	58.640	0.8502	7.3734	8.1688	5.5945
17	0.471	59.574	1.5701	4.0685	5.0728	2.7493
18	0.350	53.180	5.4599	121.7731	123.7536	126.8210
19	0.565	60.901	0.5035	1.9377	1.9621	1.1879
20	0.664	60.180	1.4118	7.2860	11.3355	20.7586
21	0.600	60.493	0.1057	1.6383	2.5304	3.6661
22	0.586	58.370	1.0703	7.8940	8.3402	27.0517
23	0.567	60.216	0.0813	0.1938	0.3622	1.1943
24	0.980	66.080	13.4982	168.1998	210.0040	133.3725
25	0.485	59.500	1.0472	2.9802	3.5583	1.6948
26	0.573	60.052	0.0214	0.2364	0.5199	2.4067
27	0.520	59.501	0.2852	1.0646	1.0547	1.3111
28	0.410	50.470	15.5514	218.4728	211.2181	310.3428
29	0.539	58.666	0.1453	4.2697	4.0543	10.7904
30	0.554	60.239	0.2011	0.1357	0.1501	0.3682

The bold numbers indicate outlying points

were modified to (0.880, 65.230) and (0.980, 66.080), respectively. The resulting Hotelling's-T² statistics for the four methods together with the data are presented in Table 6. The corresponding multivariate control charts are as shown in Figures 2a through 2d for Classical, MVE, MCD and PRM, respectively. From Table 2, while the Classical, MCD and MVE control charts indicated two points as outliers, the proposed robust method (PRM) chart signals all the three observations (points 2, 14 and 24) as outliers. The classical chart indicated observations 2 and 24 as outlying points while both MCD and MVE indicated observations 14 and 24 as outliers. Figures 3a-3d gave clearer details.

The number of outliers was increased to 5 by modifying observations 18 and 28 to (0.350, 53.180) and (0.410, 50.470), respectively in addition to the three existing outlying points in the data set. The multivariate statistics obtained from the four methods together with the data set

are shown in Table 7. From the table, the two robust methods, MCD and MVE, identified all the outliers (observations 14, 18, 24 and 28) except the second observation as outliers. The classical chart identified only two points, observations 24 and 28 as outlying points, while the proposed robust chart identified all the outlying points as outliers. Figures 4b and 4c are MVE and MCD control charts respectively showing the four observations above the upper control limits. Figures 4a and 4d are the control charts for Classical and PRM showing two and four points above the upper control limit respectively.

Finally, the number of outlying points was further increased to 7 with the modification of observations 8 and 20 to (0.400, 50.550) and (0.715, 62.455) respectively. Table 8 gave the multivariate statistics for all the four control charts. The Classical control chart's performance, as shown in Figure 5a, was so poor that it can only identified only two observations as outlying points out of

Table 8. Data set and Hotelling's-T² statistic using the Classical, MCD, MVE and the Proposed Robust Method (PRM) when there are seven (7) outliers

No.	X ₁	X ₂	Classical	MCD	MVE	Proposed
1	0.567	60.558	0.2333	0.8913	0.8913	0.6835
2	0.538	56.303	1.8549	30.6481	30.6481	67.2629
3	0.530	59.524	0.1823	0.6299	0.6299	1.6341
4	0.562	61.102	0.6502	2.6878	2.6878	2.0714
5	0.483	59.834	1.4218	3.4235	3.4235	2.0271
6	0.525	60.228	0.6607	0.5959	0.5959	0.1661
7	0.556	60.756	0.4854	1.2566	1.2566	0.7574
8	0.400	50.550	9.3145	199.6315	199.6315	298.0955
9	0.547	60.153	0.2491	0.0507	0.0507	0.2316
10	0.531	60.640	0.8746	1.2905	1.2905	1.0863
11	0.581	59.785	0.0369	1.2776	1.2776	5.2204
12	0.585	59.675	0.0717	1.7524	1.7524	6.8552
13	0.540	60.489	0.5627	0.6248	0.6248	0.2732
14	0.880	65.230	7.2106	131.8779	131.8779	83.6520
15	0.554	59.788	0.0547	0.1806	0.1806	1.9257
16	0.469	58.640	0.8629	7.7875	7.7875	5.5945
17	0.471	59.574	1.5499	4.8920	4.8920	2.7493
18	0.350	53.180	3.7306	118.1500	118.1500	126.8210
19	0.565	60.901	0.4548	1.8720	1.8720	1.1879
20	0.715	62.455	1.6623	32.0042	32.0042	21.5871
21	0.600	60.493	0.1221	2.6479	2.6479	3.6661
22	0.586	58.370	0.7348	8.4610	8.4610	27.0517
23	0.567	60.216	0.0983	0.4057	0.4057	1.1943
24	0.980	66.080	13.3108	206.7971	206.7971	133.3725
25	0.485	59.500	1.0515	3.3899	3.3899	1.6948
26	0.573	60.052	0.0389	0.6052	0.6052	2.4067
27	0.520	59.501	0.3053	0.9708	0.9708	1.3111
28	0.410	50.470	9.9472	201.9369	201.9369	310.3428
29	0.539	58.666	0.0560	3.9447	3.9447	10.7904
30	0.554	60.239	0.2110	0.1494	0.1494	0.3682

The bold numbers indicate outlying points

seven outliers in the data. Both MVE and MCD control charts given in Figures 5b and 5c, performed better by identifying all the outlying observations except the second observation as outliers. The PRM chart performed as such by identifying all the seven outliers in the data set except the twentieth observation.

SUMMARY AND CONCLUSION

The proposed robust method empirically compete favourably well with the most widely used robust methods (MVE and MCD) in detecting outliers in the presence of multiple outliers. The proposed method is better in detecting outliers than the MVE and MCD, especially when there are fewer or single outliers in the data set. As a result of the above points, the proposed method is highly recommended when there is no information as regards the

the number of outliers in a multivariate data set.

The proposed robust method of estimating the variance-covariance matrix of multivariate data combines the efficiencies of both classical and existing robust methods (MVE and MCD) of estimation. The classical method of estimation is most efficient in multivariate analysis when there is no or only a single outlier while on the other hand the existing robust methods (MVE and MCD) are more efficient in the presence of multiple outliers in a multivariate data set.

The proposed robust method (PRM) performed better and more efficient in the two extreme cases outlined above. While existing robust methods are less efficient where there is no or only one outlier, the proposed robust method is better. Likewise when there are multiple outliers, the classical method becomes less efficient while the proposed robust method was found to be efficient.

Generally, since the information on whether a multiva-

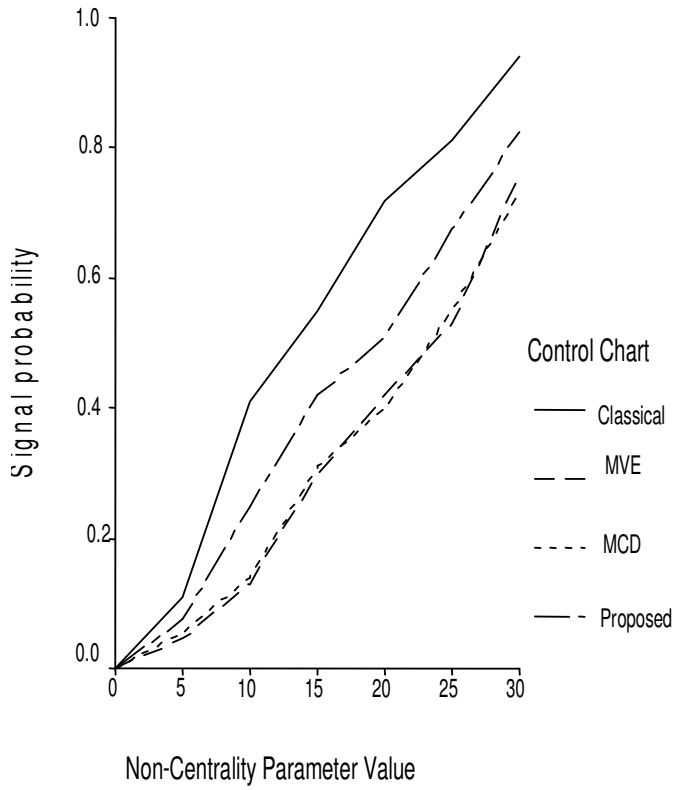


Figure 1a. Signal prob when there is one outlier.

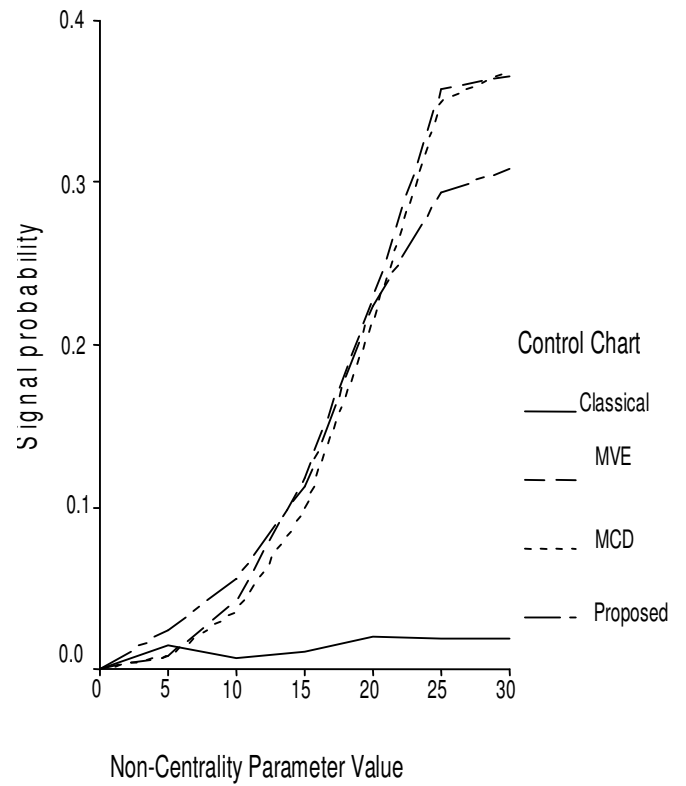


Figure 1c. Signal prob. when there are 5 outliers.

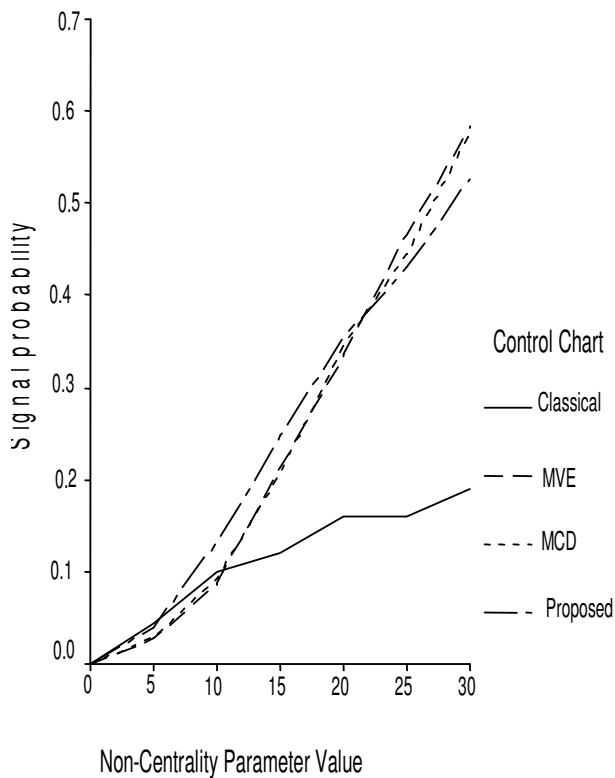


Figure 1b. Signal prob. when there are 3 outliers.

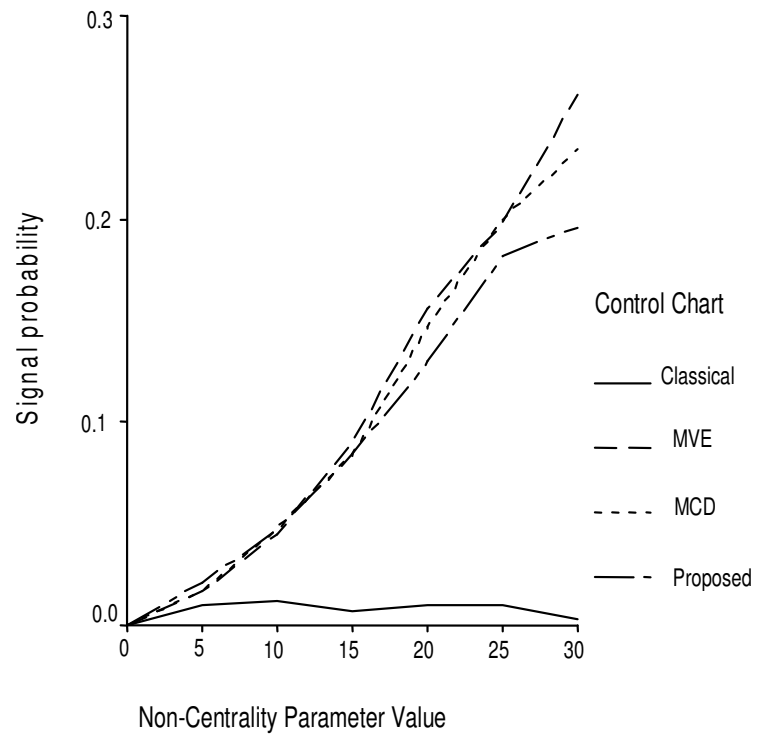


Figure 1d. Signal prob. when there are 7 outliers.

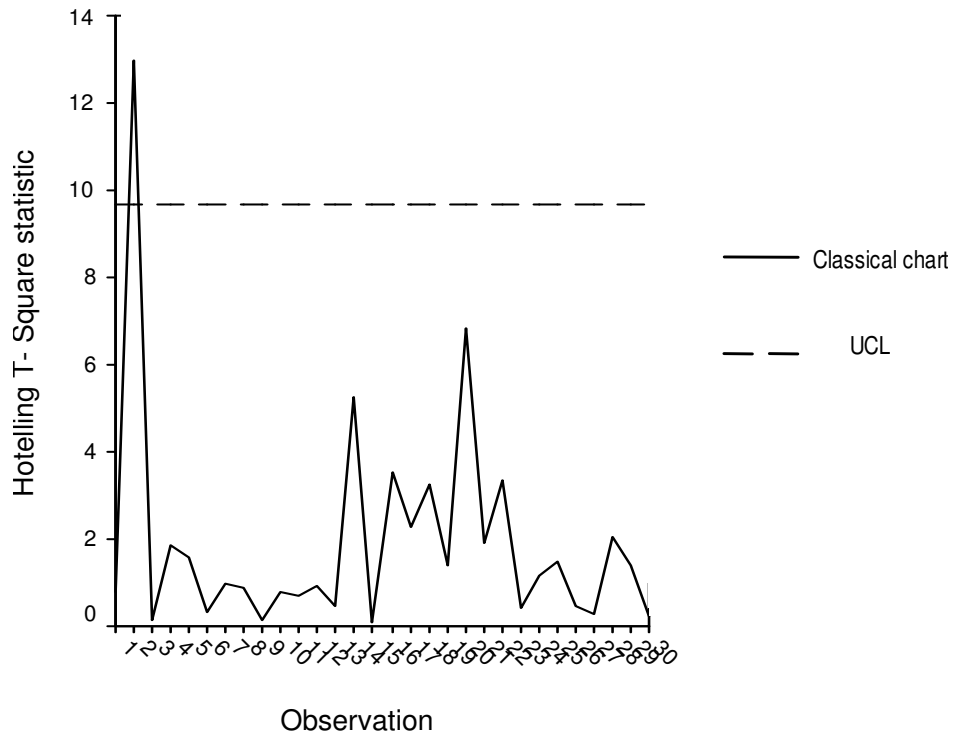


Figure 2a. Classical chart when there is one outlier

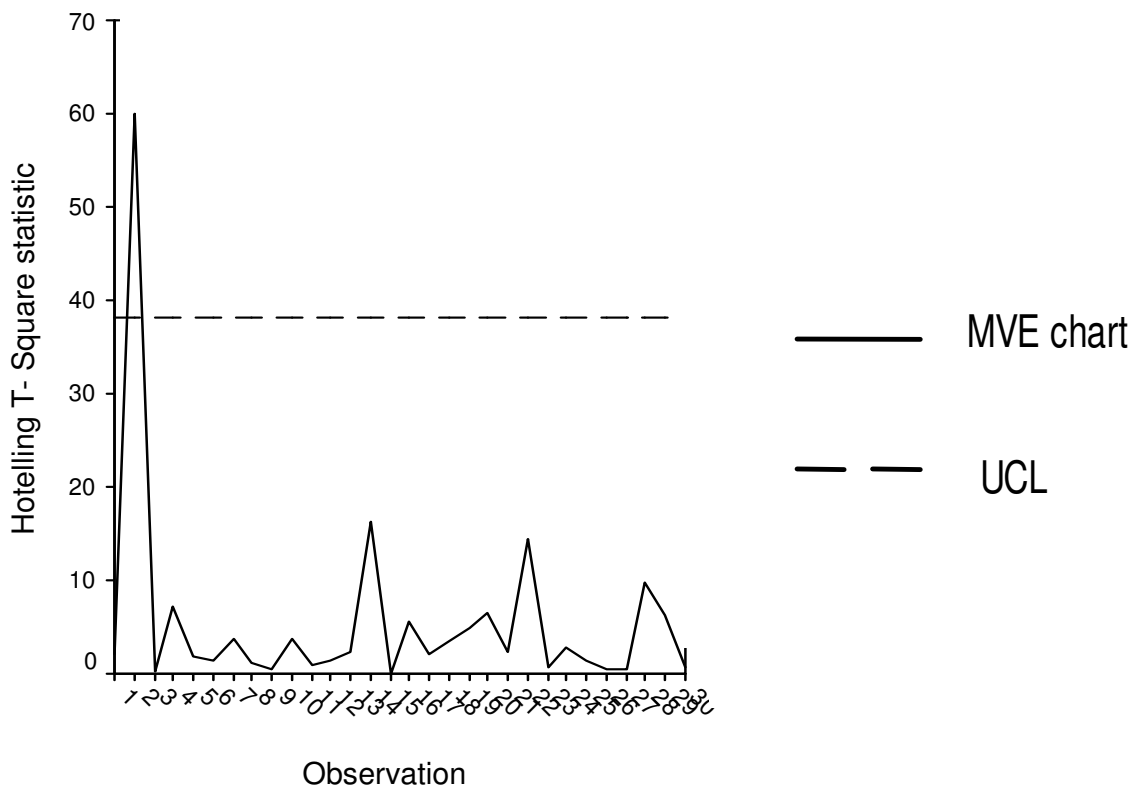


Figure 2b. MVE chart when there is one outlier.

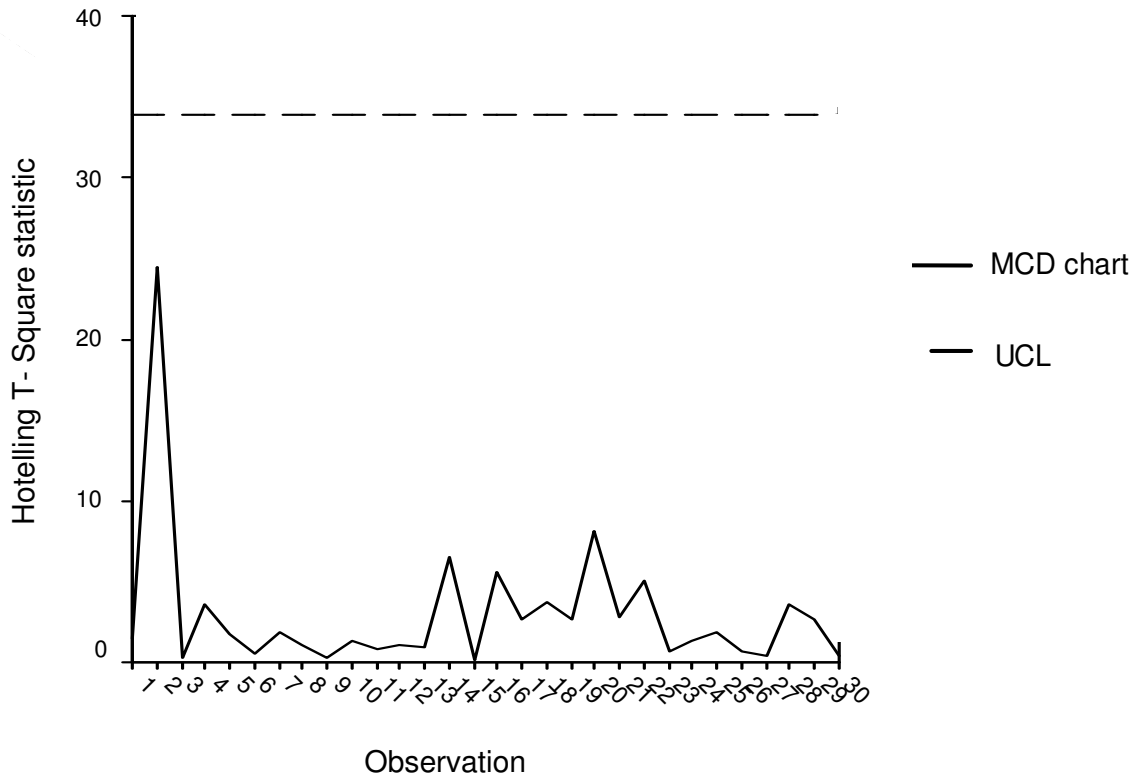


Figure 2c. MCD chart when there is one outlier.

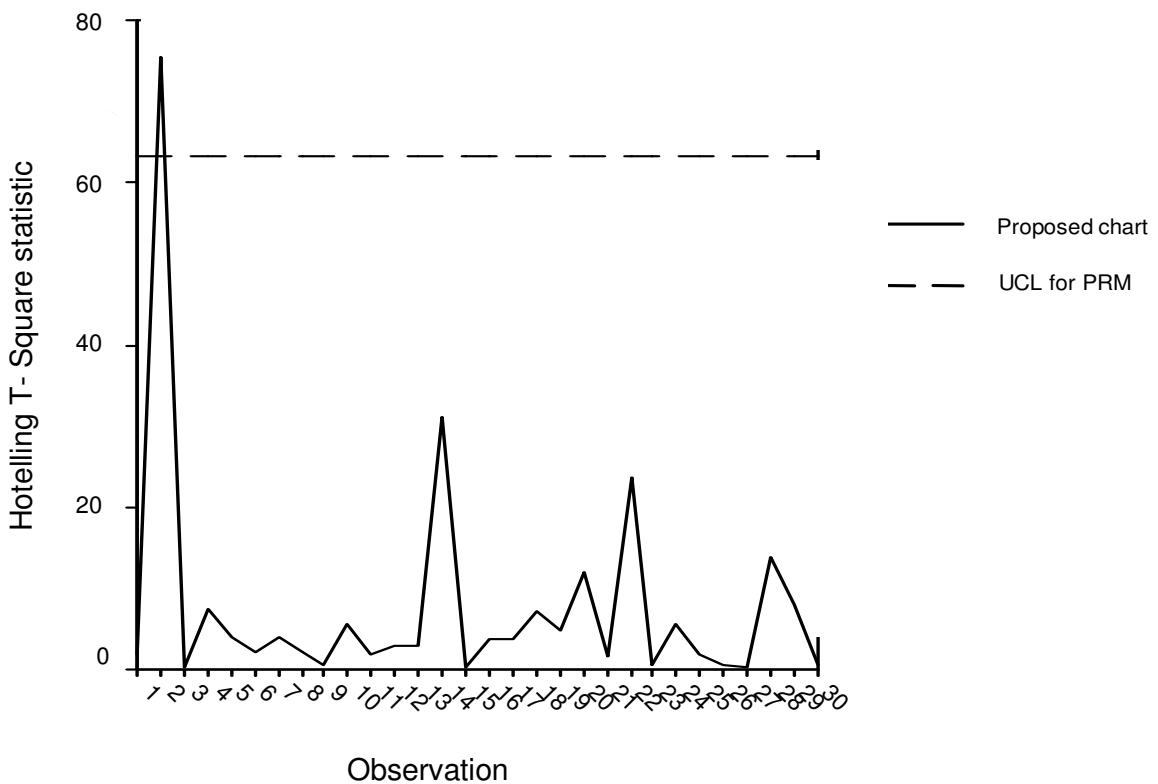


Figure 2d. PRM chart when there is one outlier

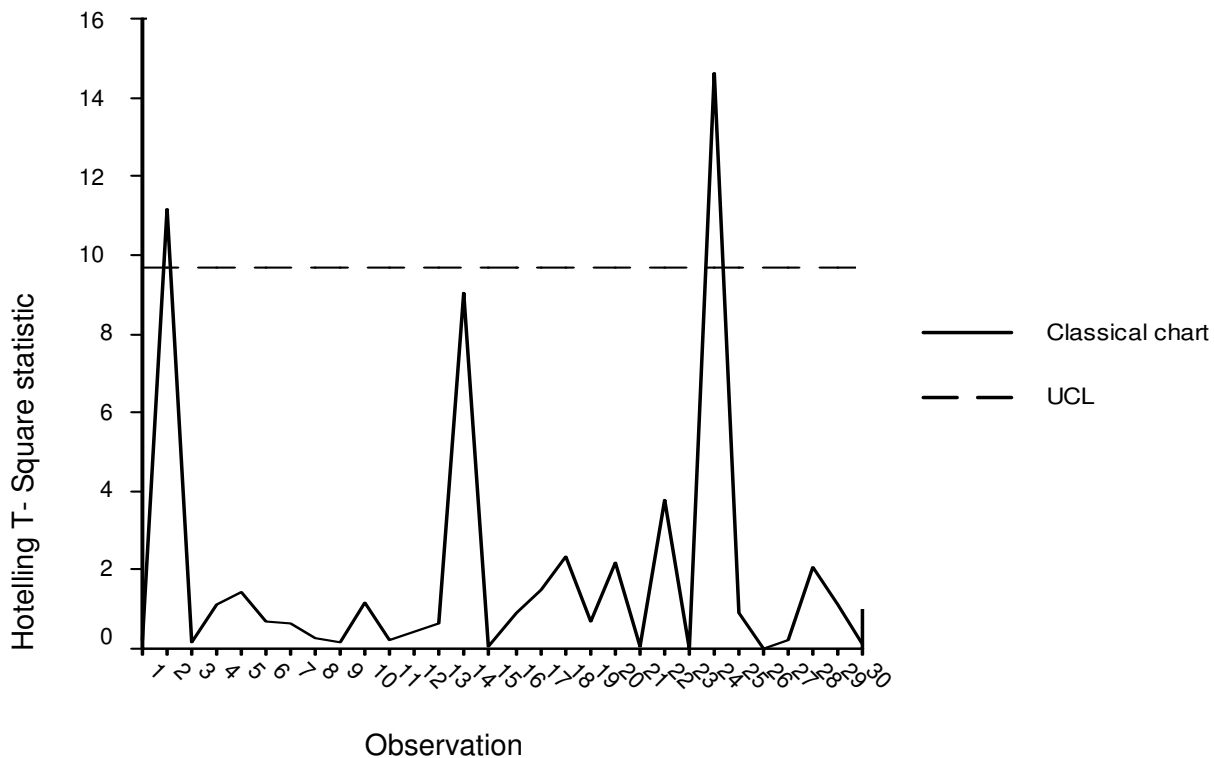


Figure 3a. Classical chart when there are 3 outliers.

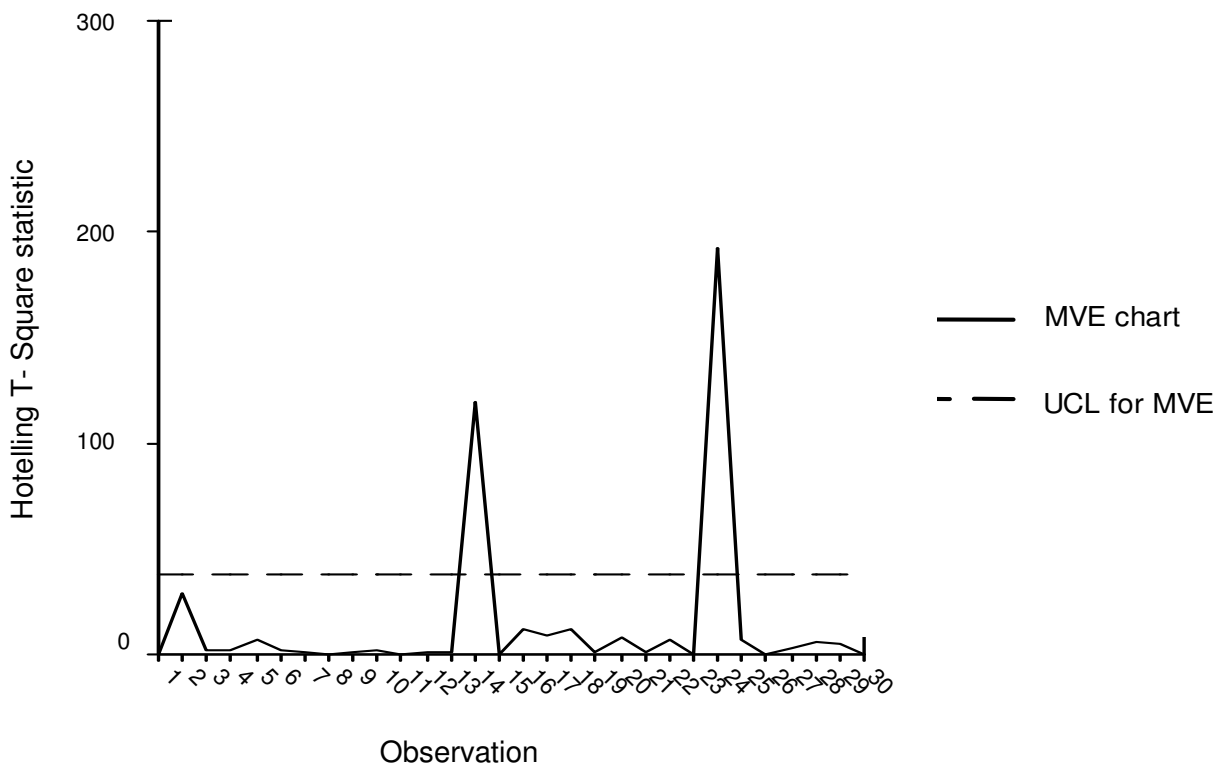


Figure 3b. MVE chart when there are 3 outliers.

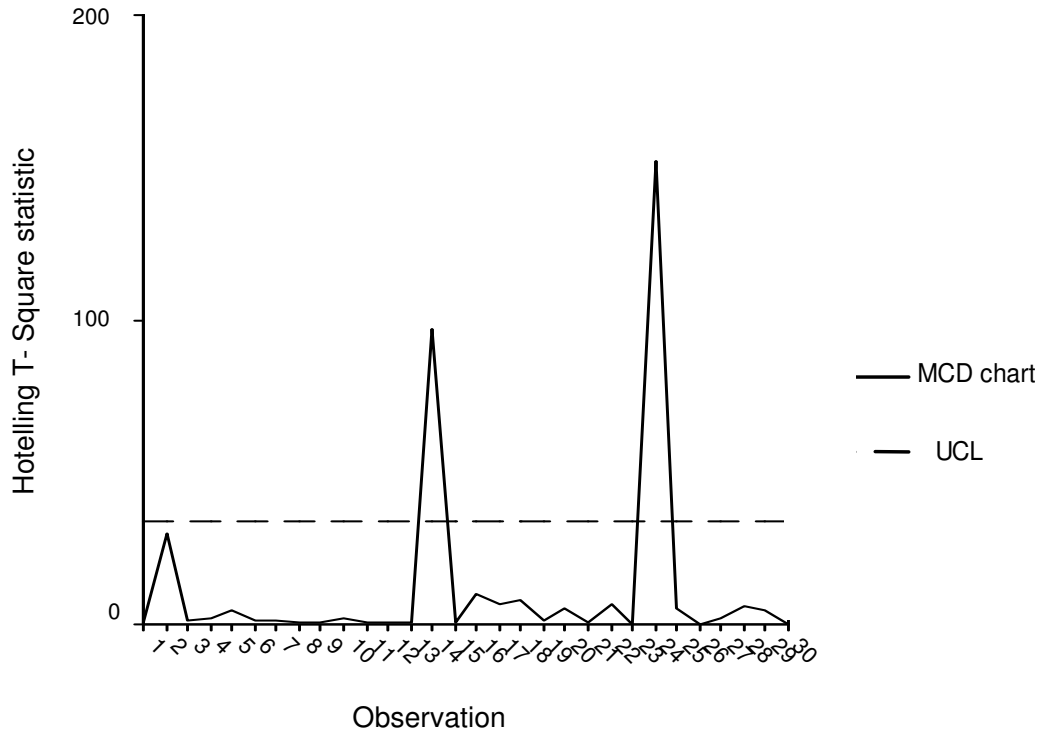


Figure 3c. MCD chart when there are 3 outliers.

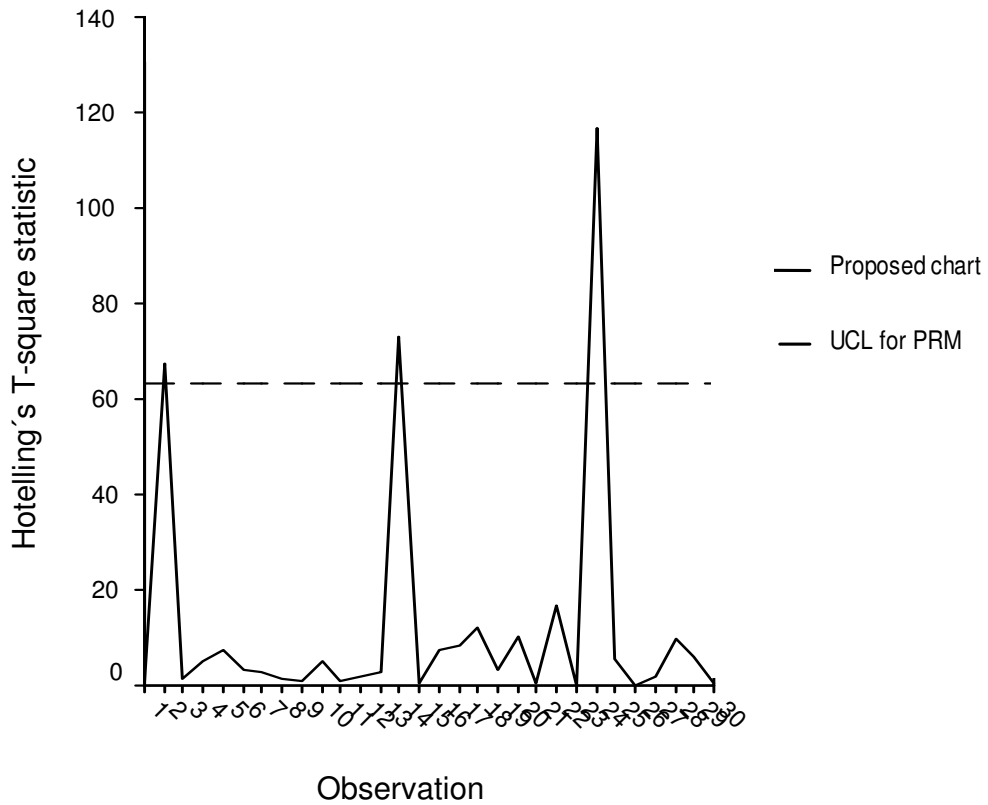


Figure 3d. PRM chart when there are 3 outliers.

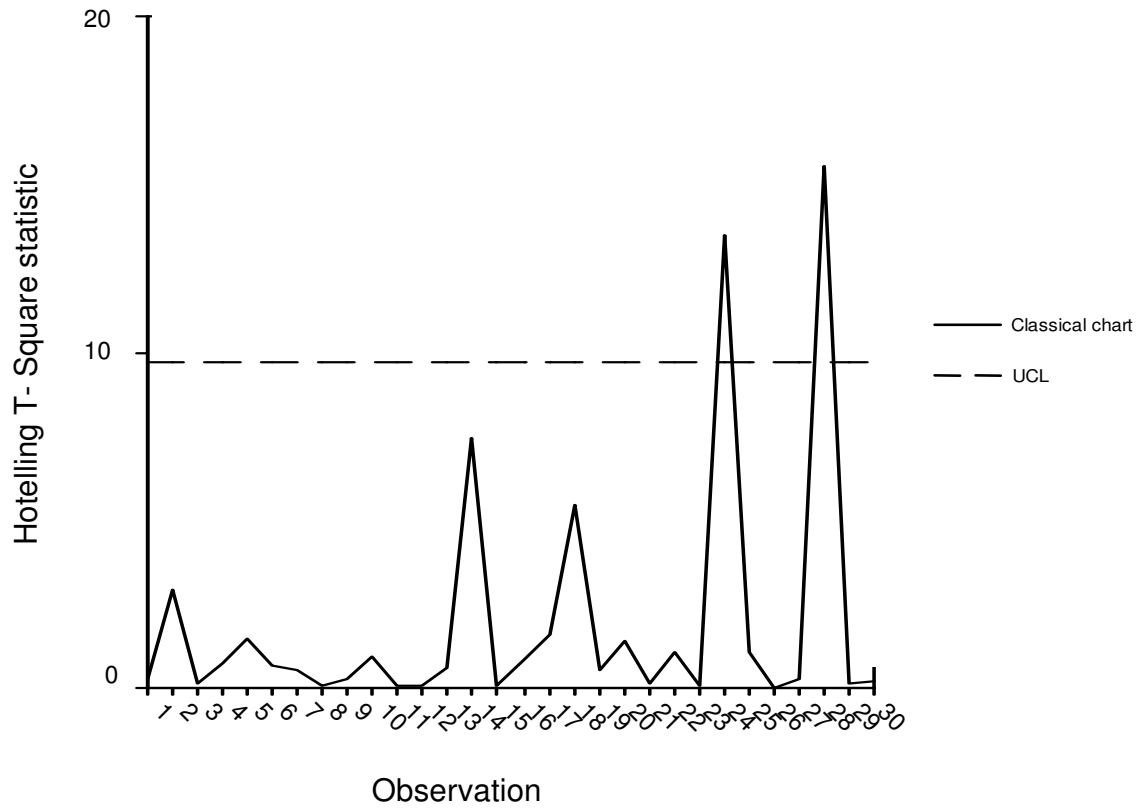


Figure 4a. Classical chart when there are 5 outliers.

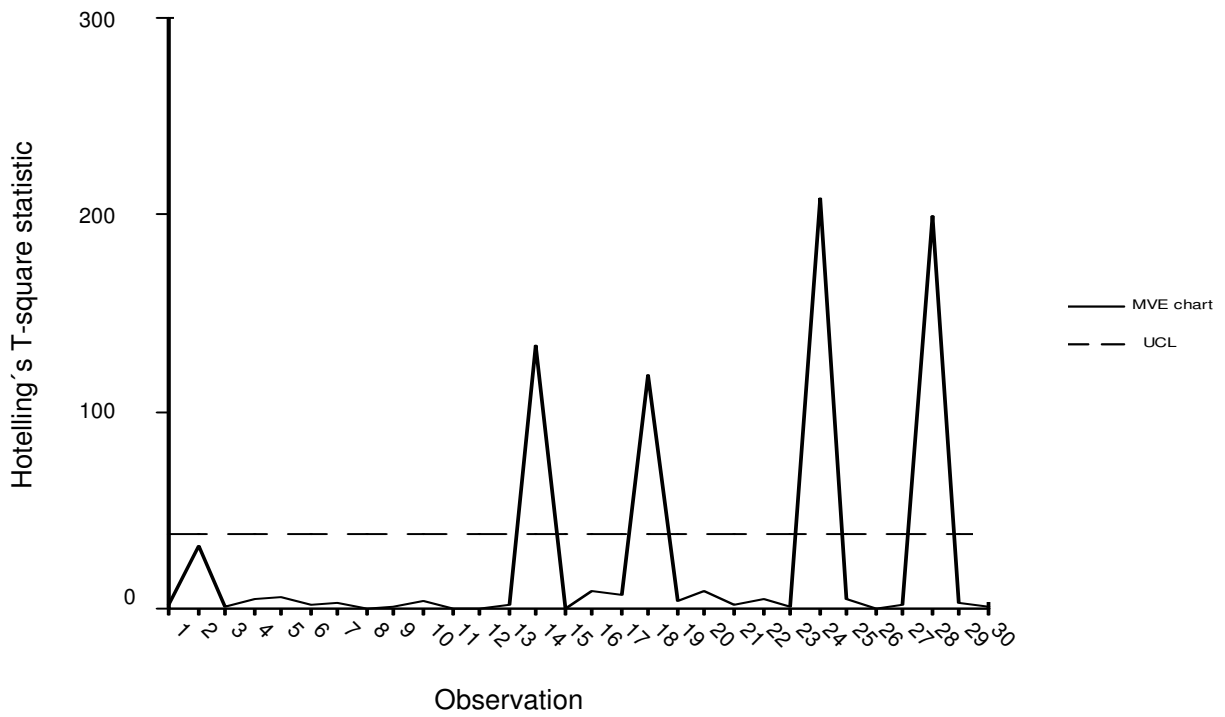


Figure 4b. MVE chart when there are 5 outliers.

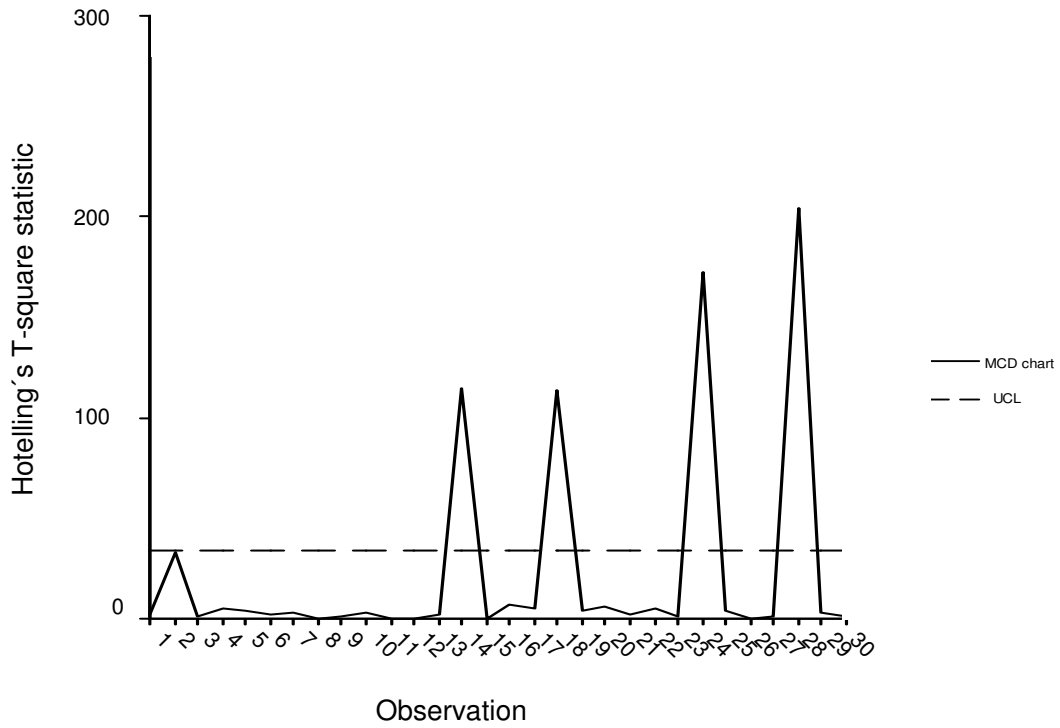


Figure 4c. MCD chart when there are 5 outliers.

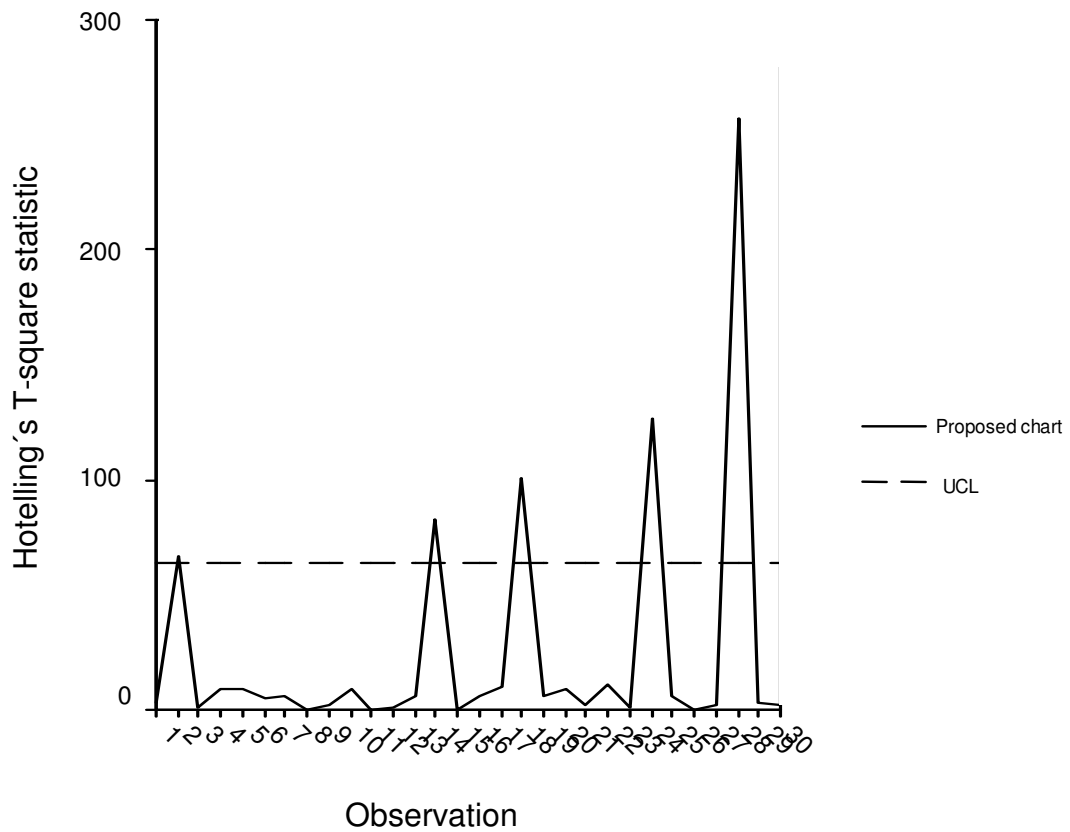


Figure 4d. PRM chart when there are 5 outliers.

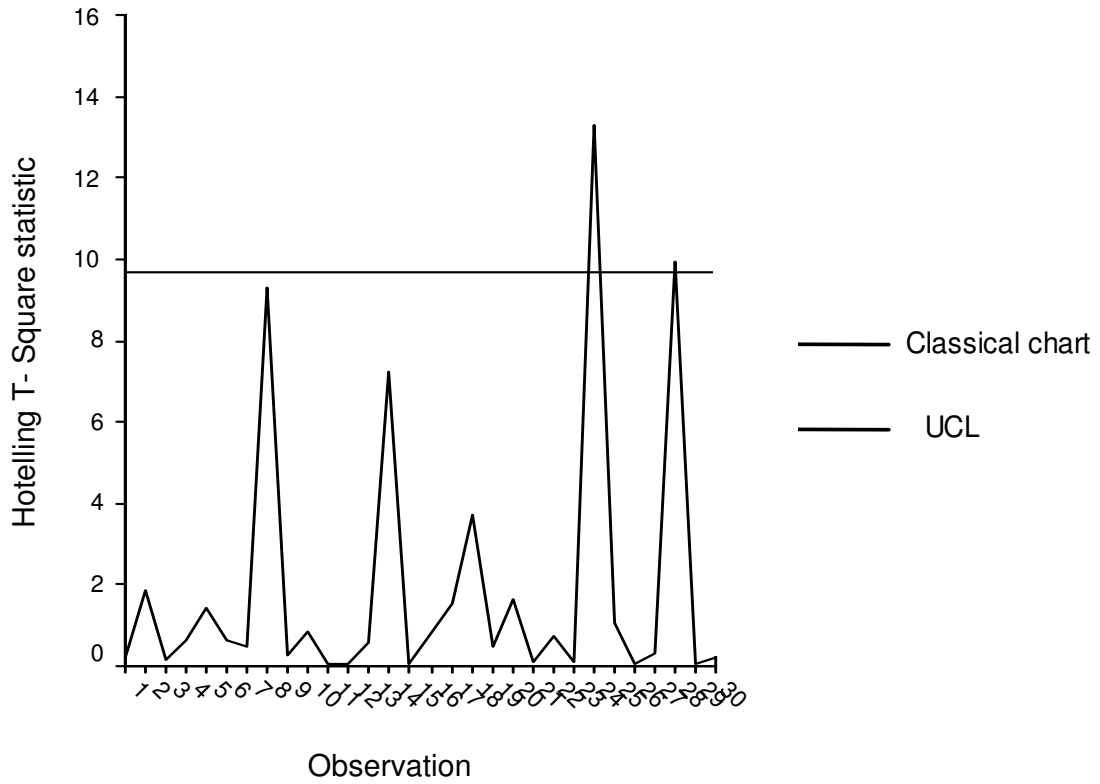


Figure 5a. Classical chart when there are 7 outliers.

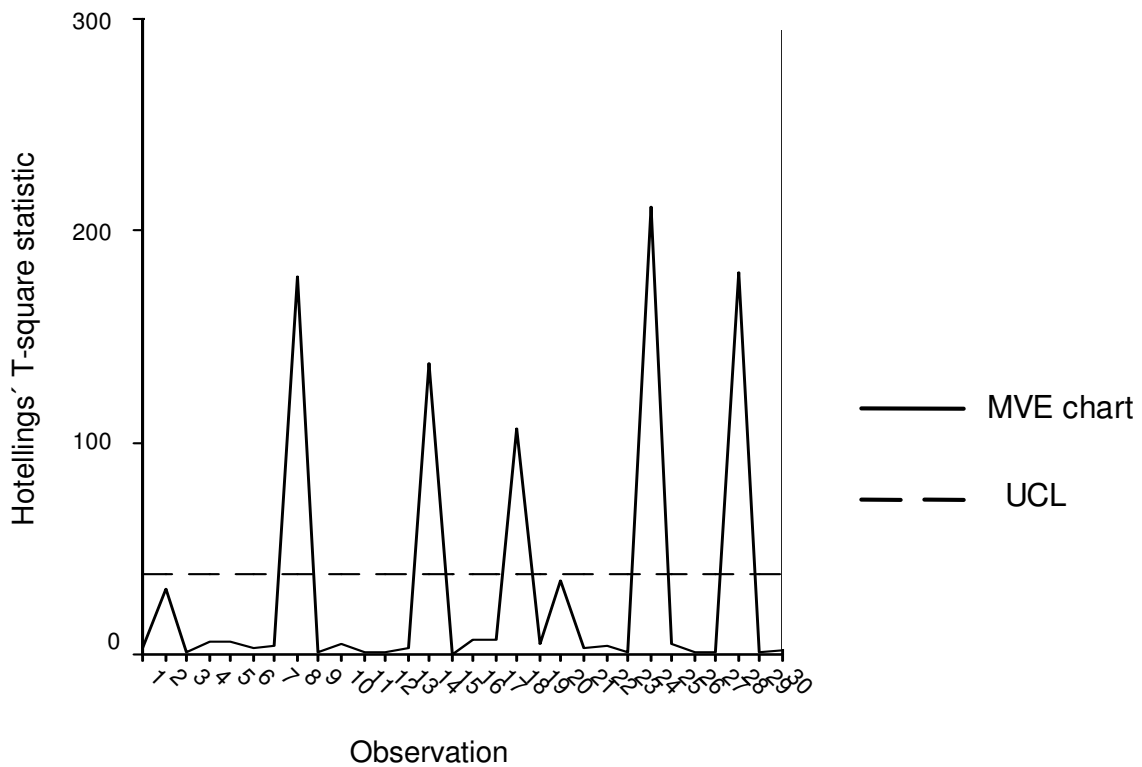


Figure 5b. MVE chart when there are 7 outliers.

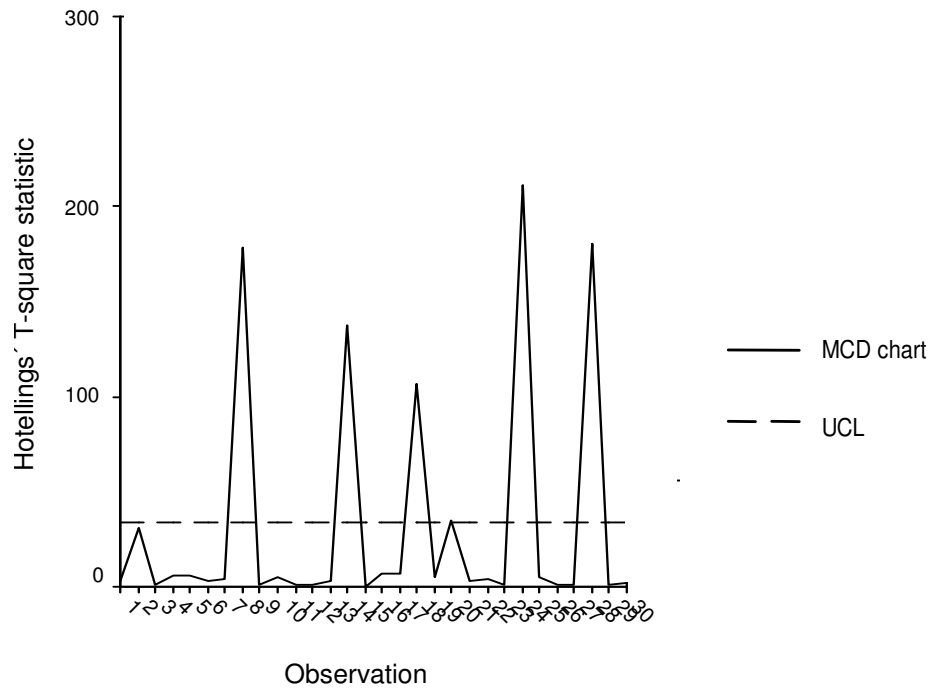


Figure 5c. MCD chart when there are 7 outliers.

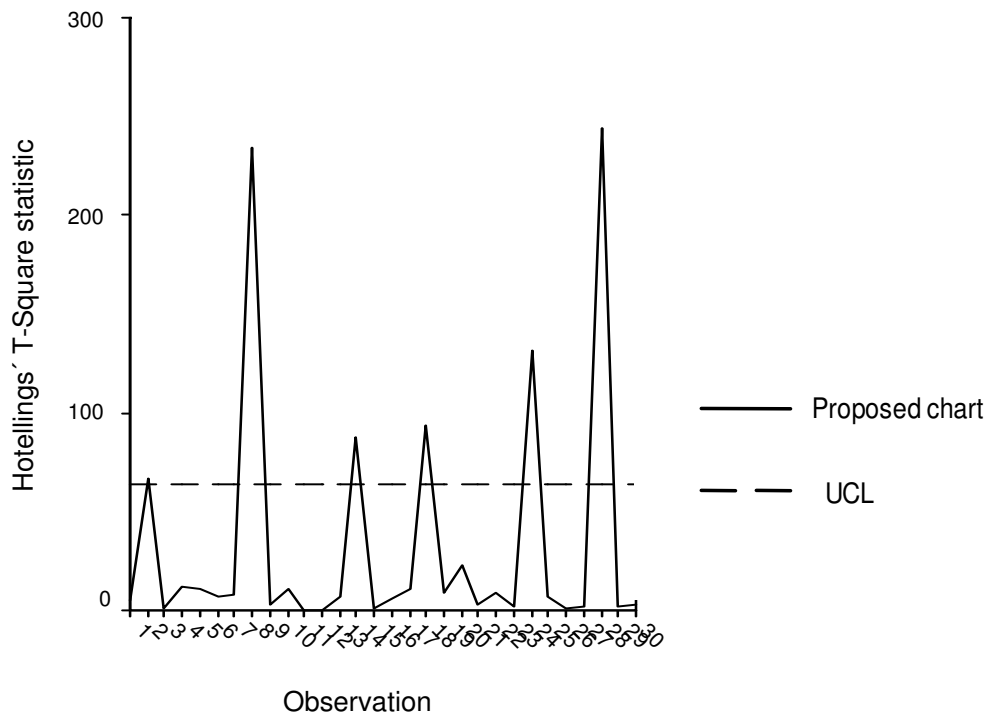


Figure 5d. PRM chart when there are 7 outliers.

iate data set contains outliers or not and even the number of outliers, may not be at the disposal of the analyst.

It is highly recommended to use the proposed robust method in estimating the variance-covariance since it will

combine both efficiencies of both classical and other robust methods in the presence or otherwise of multiple outliers. Extension of further research to analytical approach has also been opened.

REFERENCES

- Agullo J (1996). Exact Iterative Computation of the Multivariate Minimum Volume Ellipsoid with a Branch and Bound Algorithm. In Proceedings in Computational Statistics, pp.175-180.
- Daves PL (1987). Asymptotic Behavior of S-Estimates of Multivariate Location Parameters and Dispersion Matrices. *Ann. Stat.* 15: 1269-1292.
- Hubert M, Engelen S (2007). Fast Cross-validation of High-breakdown Re-sampling Methods for PCA. *Comput. Stat. Data Anal.* 51(10): 5013 – 5024.
- Jensen WA, Birch JB, Woodall WH (2002). High Breakdown Estimation Methods for Phase I Multivariate Control Charts. www.stat.vt.edu/newsletter/annual_report_statistics_2008.pdf.
- Lopuhaa HP, Rousseeuw PJ (1991). Breakdown Point of Affine Equivariant Estimators of Location and Variance Matrices. *Ann. Stat.* 27: 125 – 138.
- Vandev DL, Neykov NM (2000). Robust Maximum Likelihood in the Gaussian Case. http://www.fmi.uni-sofia.bg/fmi/statist/Personal/Vandev/papers/ascona_1992.pdf.
- Quesenberry CP (2001). The Multivariate Short-Run Snapshot Q Chart. *Qual. Eng.* 13: 679 – 683.
- Rocke DM, Woodruff DL (1998). Robust Estimation of Multivariate Location and Shape. *J. Stat. Plan. Inference* 57: 245 – 255.
- Rousseeuw PJ (1984). Least Median of Squares Regression. *J. Amer. Stat. Assoc.* 79: 871-880.
- Rousseeuw PJ, Van Zomeren BC (1990). Unmasking Multivariate Outliers and Leverage Points. *J. Am. Stat. Assoc.* 85: 633-639.
- Rousseeuw PJ, Van Zomeren BC (1991). Unmasking Multivariate Outliers and Leverage Points. Rejoinder- *J. Am. Stat. Assoc.* 85: 648-651.
- Titterton DM (1975). Optimal Design; Some Geometrical Aspect of D-Optimality. *Biometrika* 62(3): 313 – 320.
- Vargas JA (2003). Robust Estimation in Multivariate Control Charts for Individual Observations. *J. Qual. Technol.* 35(4): 367-376.