Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 11 (3): 503-511, 2011
ISSN 1812-5654 / DOI: 10.3923/jas.2011.503.511

High Breakdown Estimators to Robustify Phase II Multivariate Control Charts

[1]M. Mohammadi, [1,2]H. Midi, [1,2]J. Arasan and [3]B. Al-Talib
[1]Institute for Mathematical Research, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
[2]Department of Mathematics, Faculty of Science, Universiti Putra Malaysia,
43400 UPM Serdang, Selangor, Malaysia
[3]Department of Statistics and Informatics,
Faculty of Computer Sciences and Mathematics, University of Mosul, Mosul, Iraq

Abstract: Control chart is a statistical process control tool that is used to monitor the changes in a process. Hotelling's T^2 chart is one of the most popular control charts for monitoring independently and identically distributed random vectors. This chart detects many types of out-of-control signals, but it is not sensitive to small shifts in the mean vector. This study propose a more efficient T^2 control charts based on the re-weigted robust estimators of location and dispersion. The proposed control charts are attained by substituting the classical estimators of the mean vector and covariance matrix in the Hotelling's T^2 by the re-weighted MCD and re-weighted MVE estimators. In this study, Monte Carlo simulations were carried out to establish the proposed robust control limit. Following that, we suggested suitable estimators for each condition. Our advice in this study is replacing the classical mean vector and covariance matrix of the data in the Hotelling's T^2 statistic by there weighted MCD and Re-weighted MVE estimators.

Key words: Hotelling's T^2 chart, breakdown point, RMCD estimators, RMVE estimators, outliers

INTRODUCTION

One of the most popular tools used to construct a multivariate control chart using individual observations is the Hotelling's T^2 control chart. The main aim of a multivariate control chart is detecting the cause of the process change. In the literature, construction of this control chart is carried out in two distinct phases. In phase I operation, Hotelling's T^2 chart is often used to purge outliers whereas signal detection is done in the phase II operation. In the first phase, Historical Data Set (HDS) is used to establish the control limits in order to detect outliers. If one or more of these initial observations in HDS is out of control, the phase I control limits are recomputed based on the remaining observations. This step is repeated until the process comes to the state of control and the in-control parameters of the process are estimated. Then clean, phase I data is used to set control limits for monitoring future (phase II) data. Generally, for a retrospective phase I analysis of a Historical Data Set (HDS) our objective is defined in two steps.

In the first step, mean shifts that might result to ruins the estimation of the in-control mean vector and variance-covariance matrix are discovered. In the second step, after identification and elimination of multivariate outliers, the in-control parameters estimates which are established in the control subset of the HDS, to be used.

In the computation of the Hotelling's T^2 the classical mean and covariance are utilized, hence the results can be heavily influenced by outliers. In recent years, several multivariate statistical process control techniques have been proposed to analyze and monitor multivariate data (Tracy *et al*., 1992; Sullivan and Woodall, 1996; Vargas, 2003). In this study, in order to decrease the impact of outliers we propose the use of the re-weighted robust estimator instead of the classical estimator in the computation of Hotelling's T^2.

HOTELLING'S T^2

Hotelling *et al*. (1974) introduced a statistic that combines information from the dispersion and mean of several variables. The T^2 statistic may be computed using a single observation from p components, or it may be computed using the mean from a sample of size n. In this study, a subgroup of size 1 (i.e., a single observation) will be assumed for the T^2 computations. The T^2 statistic for a $p\times1$ multivariate normal vector, $X = (X_1,..., X_p)$ is defined as:

Corresponding Author: Habshah Midi, Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, 43400 UPM Serdang, Selangor, Malaysia Tel: +603-8946 6822 Fax: +603- 8943-7958

$$T^2 = (X_i - \bar{X})' S^{-1} (X_i - \bar{X}) \quad (1)$$

Where:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$S = \frac{1}{n-1}\sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})'$$

are the common estimators of the mean vector and covariance matrix obtained from the historical data set. Usually, we assume that the X_i's are independent multivariate normal, $MVN_p(\mu, \Sigma)$ with μ and covariance matrix Σ. In order to detect any possible signal, for each individual observation, we compare the T^2 with control limits.

Hotelling's T^2 statistic can be described by different distribution based on different situations. Suppose the parameters of the MVN distribution, μ and Σ are known. The T^2 statistic for an individual observation vector X follows the chi-square distribution:

$$T^2 = (X - \mu)' \Sigma^{-1} (X - \mu) \sim \chi^2_p \quad (2)$$

Assuming the parameters of the MVN distribution are unknown, the estimators $\bar{X}$ and S must be used. These values are calculated from HDS, consisting of n observations. The distribution of the T^2 statistic for an individual observation vector X, independent of $\bar{X}$ and S is:

$$T^2 = (X - \bar{X}) S^{-1} (X - \bar{X}) \sim \left[\frac{p(n+1)(n-1)}{n(n-p)}\right] F_{(p,n-p)} \quad (3)$$

where, F is distributed as F distribution with p and (n-p) degrees of freedom. In another case, let us say the observation vector X is not independent of the estimators $\bar{X}$ and S but is included in their computation. Then the distribution of the T^2 statistic is given as follows (Tracy *et al*., 1992):

$$T^2 = (X - \bar{X})' S^{-1} (X - \bar{X}) \sim \left[\frac{(n-1)^2}{n}\right] B_{(p/2,(n-p-1)/2)} \quad (4)$$

where, B is distributed as Beta distribution with parameters p/2 and (n-p-1)/2. In the computation of Hotelling's T^2 the classical estimators of location and dispersion have been used to estimate the population mean vector and covariance matrix. However, this classical estimators are affected by outliers in the phase I data. Therefore, we construct an alternative estimator based on Re-weighted Minimum Covariance Determinant (RMCD) and Re-weighted Minimum Volume Ellipsoid Estimators (RMVE) which have a higher efficiency and reduce the influence of outlying observations. Vargas (2003) and Jensen *et al*. (2007) recommended using the Minimum Covariance Determinants (MCD) and Minimum Volume Ellipsoid (MVE) estimators of mean vector and covariance matrix in the Hotelling's T^2 charts (Wisnowski *et al*., 2002; Chenouri *et al*., 2009). The exact distribution of the T^2 based on the RMCD and RMVE are not tractable, so we used Monte Carlo method to obtain the appropriate control limits.

ROBUST ESTIMATORS

Modified version of the MCD and MVE estimators are RMCD and RMVE. Reweighting step that is included in the computation of RMCD and RMVE, greatly increase the finite-sample efficiency of these estimators.

RMCD estimators inherit the good properties of MCD estimators such as affine equivariance, the maximal asymptotic breakdown point and asymptotic normality. In addition, a fast and efficient approximate algorithm to compute Re-weighted Minimum Covariance Determinant (RMCD) is available. The above mentioned estimators, have been studied by several authors (Rousseeuw and Van Zomeren, 1990; Lopuhaa and Rousseeuw, 1991; Willems *et al*., 2002). In contrast, the re-weighted MVE estimators of mean vector and covariance matrix with ordinary MVE, re-weighted MVE estimators are more efficient, akin to the re-weighted MCDs. So, we expect that the new robust control charts (re-weighted MVE) outperforms those based on the ordinary MVEs.

RMCD ESTIMATOR

Rousseeuw (1985) introduced the Minimum Covariance Determinant (MCD) estimator that has finite sample and asymptotic breakdown points 1/2, which is based on the subset of $h = \lfloor n\gamma \rfloor$ (where $0.5<\gamma<1$) data points whose covariance matrix has the smallest possible determinant, where $1-\gamma$ is the asymptotic breakdown point. The MCD location estimate $\bar{X}_{MCD}$ is defined as the average of these h data points and the MCD scatter estimate is given by $S_{MCD} = aC_{MCD}$ where, C_{MCD} is the covariance matrix of the subset of h points and a is the multiple of the consistency and the finite sample correction factors (Willems *et al*., 2002; Chenouri *et al*., 2009). The MCD estimators have the highest finite sample

breakdown point when h = ⌊(n+p+1)/2⌋ (Rousseeuw and Leroy, 1987). In order to estimate the RMCD estimator, for each individual observation we computed the weight based on the robust distance:

$$D(X_i)^2 = (X_i - \bar{X}_{MDC})' S_{MCD}^{-1} (X_i - \bar{X}_{MCD}) \quad (5)$$

The weight (w_t) equals 1 when the squared robust distance $D(X_i)^2$ is smaller than the cutoff value $\chi^2_{p,\eta}$ (we use the value $\eta = 0.975$ which was advocated and used by Rousseeuw and Van Driessen (1999). Otherwise the weight equals zero. Afterward we can compute the $\bar{X}_{RMCD}$ and S_{RMCD}:

$$\bar{X}_{RMCD} = \frac{\sum_{i=1}^{n} w_i X_i}{\sum_{i=1}^{n} w_i} \quad (6)$$

$$S_{RMCD} = c_{\eta,p} d_{\gamma,\eta}^{n,p} \frac{\sum_{i=1}^{n} w_i (X_i - \bar{X}_{RMCD})(X_i - \bar{X}_{RMCD})'}{\sum_{i=1}^{n} w_i} \quad (7)$$

Here, $c_{\eta,p} = \eta/p\ (\chi^2_{(p+2)} \leq q_\eta)$, makes S_{RMCD} consistent under the multivariate normal distribution and yields more reliable outlier identification. The factor $d_{\gamma,\eta}^{n,p}$ is a finite sample correction given by Pison *et al.* (2002). The most commonly used algorithm for computational purposes is FAST-MCD algorithm (Rousseeuw and Van Driessen, 1999) which has been used in this study.

RMVE ESTIMATOR

In addition to the MCD, Rousseeuw (1985) introduced the minimum volume ellipsoid (MVE) as affine equivariant and high breakdown estimators of location and dispersion. The MVE estimators are based on the smallest volume ellipsoid that covers at least h = ⌊nγ⌋ (where $0.5<\gamma<1$) observations. The MVE location estimator $t \in R_p$ and dispersion estimator C minimize the determinant of positive definite symmetric matrix C of size p, subject to the condition:

$$\#\{i : (X_i - t)' C^{-1} (X_i - t) \leq c^2\} \geq h \quad (8)$$

The constant c determines the magnitude of C and is usually chosen to make C, a consistent estimator of the covariance matrix under multivariate normal model, i.e., $c = \sqrt{\chi^2_{p,\eta}}$. Davies (1992) has shown that the MVE location estimator has a slow $n^{-1/3}$ rate of convergence and a non-normal asymptotic distribution. This low rate of convergence implies that the asymptotic efficiency of the MVE estimators is 0%. Therefore, to increase the efficiency of MVE, we propose to employ the re-weighted MVE (RMVE) similarly with the re-weighted MCD. For more details on MVE and RMVE estimators refer to the Van Aelst and Rousseeuw (2009).

DESIGNING THE ROBUST T^2 CONTROL LIMITS

As mentioned before, in phase I analysis the usual control charts use standard mean and covariance matrix estimators, which are sensitive to outliers. It is known that the Mahalanobis distance based on the classical estimators is effective in detection of a single outlier, but is not appropriate in the case of multiple outliers.

This is due to the fact that, outliers mask each other and estimator fails to detect them, which is described as the masking effect. In the literature, a variety of robust estimators have been proposed to overcome this problem and minimize the impact of outliers. To this end, we constructed a new statistic by substituting the classical estimators in the Hotelling's T^2 by the reweighted MCD and reweighted MVE estimators.

Consider that the phase I historical data set consists of n time-ordered vectors that are independent of each other. Each vector is of dimension p and there are no subsamples in the observations, the re-weighted MCD based Hotelling T^2 statistic for phase II observation $X_f \notin \{X_1, ..., X_n\}$ is define as follows:

$$T^2_{RMCD}(f) = (X_f - \bar{X}_{RMCD})' S^{-1}_{RMCD} (X_f - \bar{X}_{RMCD}) \quad (9)$$

The finite-sample distributions of the T^2_{RMCD} (f) is unknown, therefore, we computed the control limits based on the empirical distributions of respective robust T^2 charts. Several studies have been investigated the asymptotic properties of these estimators (Lopuhaa and Rousseeuw, 1991; Butler *et al.*, 1993; Croux and Haesbroec, 1999). It should be noted that the $\bar{X}_{RMCD}$ and S_{RMCD} are a good approximation to the parameters μ and Σ and $X_f \sim MVN_p(\mu, \Sigma)$. In this case, if we use the Slutsky theorem, as $n \to \infty$ the asymptotic distribution of robust T^2 is χ^2_p. This asymptotic distribution is only applicable when n is large. In the case of small sample size, we apply Monte Carlo simulations to estimate the quantiles of the $T^2_{RMCD}(f)$, for several combinations of sample sizes and dimensions. The robust T^2 statistic based on the reweighted MVE estimators of location and dispersion is defined as:

$$T^2_{RMVE}(f) = (X_f - \bar{X}_{RMVE})' S^{-1}_{RMVE} (X_f - \bar{X}_{RMVE}) \quad (10)$$

where, $\bar{X}_{RMVE}$ and S_{RMVE} are the re-weighted MVE estimates of mean vector and covariance matrix of phase I process, respectively. Since, the distribution or even asymptotic distribution of $T^2_{RMCD}(f)$ is not known, so, we obtain the quantiles of the $T^2_{RMVE}(f)$ for different combination of dimensions and sample sizes.

SIMULATION

We generated 5000 samples of size n from a p-multivariate standard normal distribution $MVN_p(0, I_p)$. Our simulation process in phase I and phase II was constructed as follows:

Phase I: We computed the re-weighted MCD and MVE mean vector and covariance matrix estimates ($\bar{X}_{RMCD}$, $\bar{X}_{RMVE}$ and S_{RMCD}, S_{RMVE}) for each data set of size n.

Phase II: In addition, as phase II observation, for each data set we randomly generated a new observation X_f from $MVN_p(0, I_p)$ and calculated the respective T^2_{RMCD} and T^2_{RMVE} value as given by Eq. 9, 10. Scatter plots of the empirical 99% quantiles of T^2_{RMCD} and T^2_{RMVE} versus the sample size n for different dimensions are presented in Fig. 1 and 2.

Since, the creation of control limits for each sample size is weariful and time consuming, it would be preferable to have a formula for the calculating of control limits. Then we sketched the graph of the simulated quantiles for the $T^2_{RMCD}(f)$ and $T^2_{RMVE}(f)$ estimators versus the size of the data set n. A regression curve was fitted to smoothly predict the control limits for any phase I sample size n, dimensions p = 2, 6, 10 and confidence level $1-\alpha = 0.99$. With respect to the graph, we utilized the regression equation $f(n) = \beta_1+\beta_2/n\beta_3$ for the quantiles. As mentioned earlier, since the $T^2_{RMCD}(f)$ is asymptotically distributed as χ^2_p, it is reasonable to use the $\chi^2_{(p,1-\alpha)}$ instead of the $\beta_{1,p,1-\alpha,\gamma}$ in Eq. 11:

$$f_{p,1-\alpha,\gamma}(n) = \beta_{1p,1-\alpha,\gamma} + \frac{\beta_{2,p,1-\alpha,\gamma}}{n^{\beta_{3,p,1-\alpha,\gamma}}} \quad (11)$$

where, $\chi^2_{(p,1-\alpha)}$ is the 1-α quantile of the χ^2_p distribution with p degrees of freedom and $\beta_{2,p,1-\alpha,\gamma}$ and $\beta_{3,p,1-\alpha,\gamma}$ are constants. Three parameter curves give better fit for control limits of RMVE estimators based on T^2 charts.

Table 1: The least square estimates of the regression parameters $\beta_{1,p,1-\alpha,\gamma}$ and $\beta_{2,p,1-\alpha,\gamma}$ and $\beta_{3,p,1-\alpha,\gamma}$ for p = 2, 6, 10 confidence levels $1-\alpha = 0.99$ and breakdown points $\gamma = 0.5, 0.75$ for RMCD estimators

p	γ	$\beta_{1,p,1-\alpha,\gamma}$	$\beta_{2,p,1-\alpha,\gamma}$	$\beta_{3,p,1-\alpha,\gamma}$
2	0.50	9.210	1707.757	1.681
	0.75	9.210	426.870	1.421
6	0.50	16.812	1877712.000	2.922
	0.75	16.812	11343.827	1.886
10	0.50	23.209	42231550.000	3.395
	0.75	23.209	80001.314	2.119

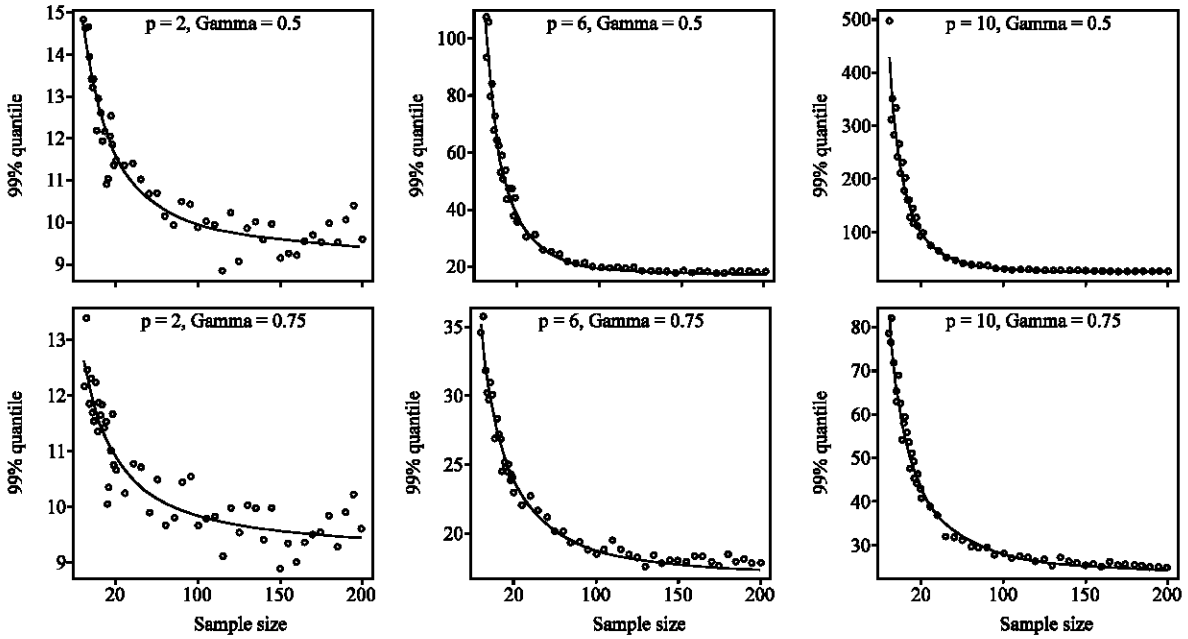

Fig. 1: The 99% simulated quantiles of T^2_{RMCD} and the fitted curves for p = 2; 6; 10, γ = 0.5 (upper panel) and γ = 0.75 (lower panel)

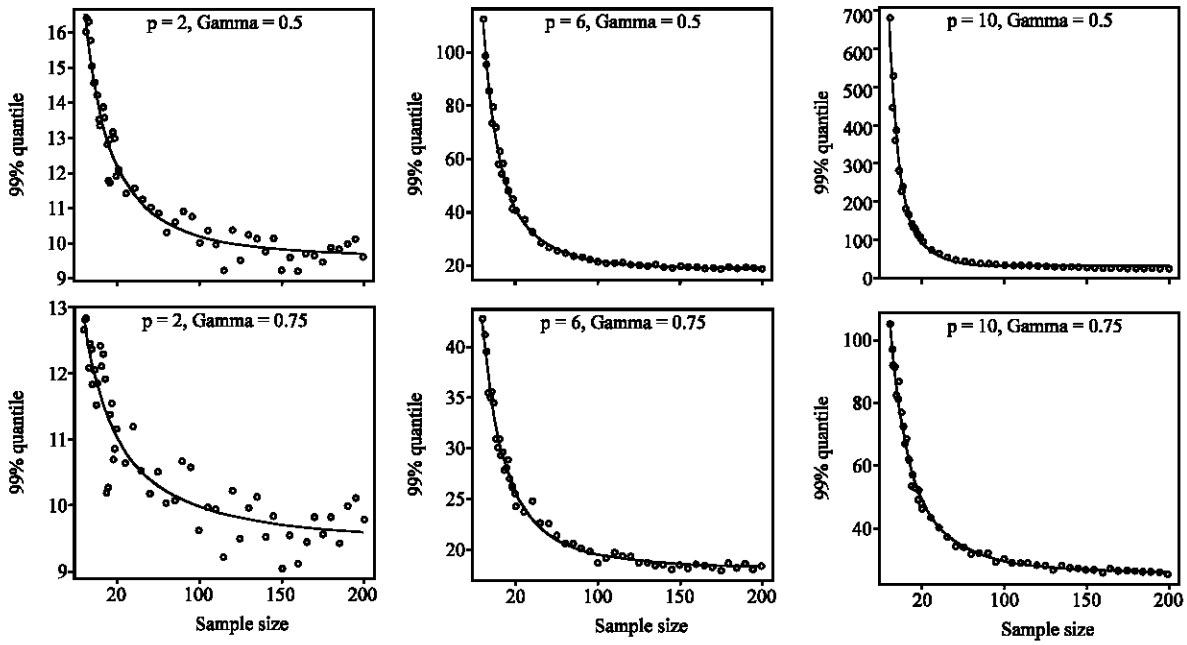

Fig. 2: The 99% simulated quantiles of T^2_{RMVE} and the fitted curves for p = 2; 6; 10, γ = 0.5 (upper panel) and γ = 0.75 (lower panel)

Table 2: The least square estimates of the regression parameters $\beta_{1,p,1-\alpha,\gamma}$ and $\beta_{2,p,1-\alpha,\gamma}$ and $\beta_{3,p,1-\alpha,\gamma}$ for p = 2, 6, 10 confidence levels 1-α = 0.99 and breakdown points γ = 0.5, 0.75 RMVE estimators

p	γ	$\beta_{1,p,1-\alpha,\gamma}$	$\beta_{2,p,1-\alpha,\gamma}$	$\beta_{3,p,1-\alpha,\gamma}$
2	0.50	9.50000	4929.218	1.925000
	0.75	9.37400	504.659	1.458000
6	0.50	17.75400	923367.400	2.708928
	0.75	18.01300	55272.395	2.275000
10	0.50	31.41108	4554712000.000	4.666000
	0.75	25.05600	321790.800	2.437000

Non linear least squares method was applied to estimate the parameters $\beta_{1,p,1-\alpha,\gamma}$, $\beta_{2,p,1-\alpha,\gamma}$ and $\beta_{3,p,1-\alpha,\gamma}$ which is shown in Table 1 and 2. Finally, in order to compute the robust control limits for p = 2, 6, 10 with varying sample size n, Table 1, 2 and Eq. 9 and 10 were used.

PERFORMANCE COMPARISON

Here, we designed a simulation study to assess the performance of our proposed methods. We considered different number of variables (p = 2, 6, 10) and the number of observations (n = 50, 150), breakdown point (γ = 0.5, 0.75), the proportion of outliers (π = 0.10, 0.20) and the amount of the shift in the process mean (δ = 3, 5). Our simulation studies include a no-outlier pattern and a pattern with multiple outliers. We can decide about the performance of the control chart by computing the probability of changes detection based on the Phase II data. The efficiency of control chart is determined by probability of signal that is the proportion of the T^2_{RMCD} and T^2_{RMVE} that is located over the control limit, using 1500 replications. We let $\delta^2 = (\mu-\mu_A)'\Sigma^{-1}(\mu-\mu_A)$ be the non centrality parameter that measures the severity of a shift of the out of control mean vector μ_A compared to the in control mean vector μ. We can without loss of generality, use a zero mean vector as μ and the identity covariance matrix as Σ, due to affine equivariance property. To simulate changes in the process mean, the following cases are considered:

- In phase I we generated clean data set (no outlier) from the standard multivariate normal distribution $MVN_p(0, I_p)$, π of them are random data points generated from the out-of-control distribution ($MVN_p(\mu_i, I_p)$) and the other 1-π observations were generated from the in-control distribution ($MVN_p(0, I_p)$)
- Phase II data were generated from $MVN_p(0, \mu_{II})$ where, $\delta^2_{II} = \|\mu_{II}\|^2$. To evaluate the performance of various methods in each simulation, we generated a sample of size n and computed the $T^2_{RMCD}(f)$ and $T^2_{RMVE}(f)$ for breakdown points of γ = 0.5, 0.75 from Eq. 9 and 10, for each observation in phase II data. Then, we compared the $T^2_{RMCD}(f)$ and $T^2_{RMVE}(f)$ to the approximate control limits

The purpose of this study is to compare the proposed method with other techniques such as standard

Table 3: Probability of signal in phase II, when phase I data is outlier free, for p =2, 6, 10 and sample size n = 50, 150 with different values of shift in phase II process mean vector μ_{II}

N	P	μ_{II}	T^2	T^2_{MCD}	T^2_{MVE}	T^2_{RMCD50}	T^2_{RMCD75}	T^2_{RMVE50}	T^2_{RMVE75}
50	2	0	0.008	0.007	0.007	0.009	0.008	0.007	0.006
		3	0.680	0.540	0.550	0.580	0.600	0.490	0.550
		5	1.000	0.840	0.870	0.870	0.900	0.840	0.890
	6	0	0.007	0.007	0.006	0.008	0.007	0.007	0.007
		3	0.520	0.320	0.330	0.330	0.370	0.300	0.340
		5	0.850	0.790	0.800	0.830	0.810	0.790	0.810
	10	0	0.004	0.004	0.005	0.004	0.005	0.004	0.004
		3	0.500	0.290	0.320	0.310	0.330	0.280	0.320
		5	0.780	0.700	0.680	0.660	0.700	0.650	0.730
150	2	0	0.001	0.005	0.003	0.004	0.006	0.005	0.007
		3	0.620	0.610	0.620	0.610	0.600	0.610	0.630
		5	0.990	0.990	0.970	0.970	0.990	0.980	0.990
	6	0	0.004	0.004	0.006	0.007	0.005	0.006	0.005
		3	0.470	0.450	0.440	0.440	0.470	0.450	0.480
		5	0.960	0.950	0.960	0.940	0.960	0.950	0.960
	10	0	0.007	0.007	0.006	0.008	0.005	0.007	0.004
		3	0.420	0.400	0.390	0.420	0.430	0.400	0.430
		5	0.950	0.950	0.930	0.940	0.950	0.940	0.950

Table 4: Probability of signal in phase II, when there is a slight shift in phase I process mean vector (μ_I = 5), for p = 2, 6, 10 and sample size n = 50 with different values of shift in phase II process mean vector (μ_{II})

N	P	μ_{II}	T^2	T^2_{MCD}	T^2_{MVE}	T^2_{RMCD50}	T^2_{RMCD75}	T^2_{RMVE50}	T^2_{RMVE75}
10%	2	0	0.006	0.005	0.006	0.008	0.007	0.007	0.006
		3	0.370	0.400	0.430	0.610	0.600	0.640	0.670
		5	0.770	0.810	0.850	0.910	0.930	0.980	0.990
	6	0	0.004	0.005	0.006	0.005	0.005	0.007	0.006
		3	0.330	0.350	0.370	0.500	0.470	0.540	0.580
		5	0.700	0.770	0.780	0.870	0.890	0.930	0.940
	10	0	0.007	0.004	0.005	0.005	0.004	0.009	0.002
		3	0.230	0.230	0.240	0.430	0.450	0.500	0.530
		5	0.480	0.570	0.620	0.840	0.870	0.880	0.900
20%	2	0	0.007	0.007	0.006	0.008	0.005	0.007	0.004
		3	0.250	0.470	0.490	0.530	0.550	0.610	0.640
		5	0.600	0.730	0.750	0.930	0.950	0.950	0.960
	6	0	0.004	0.003	0.004	0.006	0.007	0.007	0.004
		3	0.230	0.440	0.460	0.500	0.510	0.520	0.530
		5	0.570	0.630	0.670	0.910	0.940	0.940	0.950
	10	0	0.005	0.006	0.008	0.006	0.008	0.007	0.007
		3	0.050	0.260	0.250	0.500	0.470	0.500	0.510
		5	0.220	0.390	0.450	0.750	0.830	0.840	0.870

T^2 chart, robust T^2_{MCD} and T^2_{MVE} estimators discussed in Vargas (2003) and Jensen *et al.* (2007). The probability of signal on the same data sets is computed. The MCD and MVE techniques are applied in phase I to detect outliers. Then we made the standard T^2 chart based on the clean phase I data set and compared the corresponding T^2 with an appropriate quantile of F distribution to monitor phase II data. The probability of signal for detecting outliers for different phase II process shifts in the mean vector are depicted in Table 3-7

δ_I = 0 (there was no outliers in the phase I): In all cases, from Table 3-7, it is visible, by increasing the non-centrality parameter the probability of signal increases as well.

In the case of small sample sizes, we observed that the best estimator was the standard T^2, which is also supported by the previous studies (Wisnowski *et al.*, 2002; Vargas, 2003; Jensen *et al.*, 2007; Chenouri *et al.*, 2009). Conversely, as the sample size increased to 150, the performance of all robust control charts are similar to the standard T^2 chart and the performance was satisfactory.

δ_I = 5 (small proportion of outlier in phase I): As shown in Table 4 and 5, when there was a slight shift in the phase I process mean vector, for small sample sizes T^2_{RMVE} slightly outperformed the T^2_{RMCD} and they perform better than the classical T^2 and ordinary MCD and MVE. On the other hand, as the sample size increased, the performance of T^2_{RMCD} is slightly more superior to T^2_{RMVE} and other methods.

δ_I = 30 (large proportion of outlier in phase I): Based on Table 6 and 7, when the non-centrality parameter in phase I was large (δ_I = 30), the Re-weighted robust control charts for the breakdown point of γ = 0.5 and 0.75 surpassed the MCD, MVE and the classical T^2 charts. In the case of small sample size, the performance of the

Table 5: Probability of signal in phase II, when there is a slight shift in phase I process mean vector ($\mu_I = 5$), for p = 2, 6, 10 and sample size n =150 with different values of shift in phase II process mean vector (μ_{II})

N	P	μ_{II}	T^2	T^2_{MCD}	T^2_{MVE}	T^2_{RMCD50}	T^2_{RMCD75}	T^2_{RMVE50}	T^2_{RMVE75}
10%	2	0	0.003	0.004	0.004	0.002	0.002	0.004	0.002
		3	0.330	0.480	0.440	0.650	0.670	0.620	0.630
		5	0.630	0.740	0.710	0.840	0.850	0.830	0.840
	6	0	0.005	0.005	0.004	0.006	0.006	0.008	0.007
		3	0.270	0.470	0.420	0.630	0.660	0.620	0.640
		5	0.400	0.560	0.530	0.820	0.860	0.800	0.830
	10	0	0.005	0.007	0.006	0.005	0.006	0.007	0.005
		3	0.320	0.480	0.430	0.660	0.690	0.650	0.630
		5	0.600	0.760	0.700	0.850	0.870	0.810	0.840
20%	2	0	0.006	0.005	0.005	0.008	0.005	0.008	0.006
		3	0.290	0.330	0.320	0.460	0.420	0.430	0.400
		5	0.780	0.810	0.800	0.980	0.990	0.970	0.980
	6	0	0.003	0.005	0.004	0.003	0.006	0.007	0.006
		3	0.360	0.470	0.460	0.970	0.980	0.940	0.950
		5	0.730	0.820	0.840	1.000	1.000	0.980	0.990
	10	0	0.004	0.005	0.003	0.006	0.006	0.005	0.003
		3	0.670	0.680	0.680	0.960	0.980	0.950	0.960
		5	0.770	0.890	0.880	1.000	0.990	1.000	0.990

Table 6: Probability of signal in phase II, when there is a large amount of shift in phase I process mean vector ($\mu_I = 30$), for p = 2, 6, 10 and sample size n =50 with different values of shift in phase II process mean vector (μ_{II})

N	P	μ_{II}	T^2	T^2_{MCD}	T^2_{MVE}	T^2_{RMCD50}	T^2_{RMCD75}	T^2_{RMVE50}	T^2_{RMVE75}
10%	2	0	0.003	0.003	0.004	0.005	0.008	0.007	0.004
		3	0.010	0.490	0.530	0.690	0.710	0.710	0.750
		5	0.120	0.600	0.660	0.980	0.990	1.000	1.000
	6	0	0.006	0.005	0.004	0.005	0.006	0.008	0.007
		3	0.010	0.410	0.490	0.860	0.970	0.890	0.980
		5	0.070	0.580	0.630	1.000	1.000	0.980	0.990
	10	0	0.005	0.007	0.006	0.005	0.006	0.004	0.002
		3	0.060	0.320	0.380	0.730	0.970	0.910	0.990
		5	0.090	0.510	0.560	0.980	0.990	1.000	1.000
20%	2	0	0.002	0.005	0.007	0.008	0.006	0.007	0.007
		3	0.020	0.530	0.580	0.500	0.550	0.520	0.570
		5	0.030	0.690	0.730	0.920	0.960	0.940	0.990
	6	0	0.005	0.004	0.002	0.003	0.003	0.002	0.003
		3	0.010	0.490	0.520	0.820	0.950	0.830	0.970
		5	0.040	0.620	0.670	0.940	0.970	0.990	1.000
	10	0	0.007	0.007	0.005	0.007	0.008	0.006	0.000
		3	0.010	0.540	0.440	0.810	0.960	0.870	0.980
		5	0.010	0.570	0.610	0.960	0.970	0.980	0.990

Table 7: Probability of signal in phase II, when there is a large amount of shift in phase I process mean vector ($\mu_I = 30$), for p = 2, 6, 10 and sample size n =150 with different values of shift in phase II process mean vector μ_{II}

N	P	μ_{II}	T^2	T^2_{MCD}	T^2_{MVE}	T^2_{RMCD50}	T^2_{RMCD75}	T^2_{RMVE50}	T^2_{RMVE75}
10%	2	0	0.005	0.005	0.006	0.007	0.008	0.007	0.007
		3	0.050	0.410	0.360	0.800	0.790	0.770	0.780
		5	0.180	0.640	0.380	1.000	1.000	0.990	0.990
	6	0	0.004	0.006	0.006	0.007	0.006	0.007	0.005
		3	0.040	0.540	0.510	0.980	0.980	0.960	0.970
		5	0.100	0.650	0.590	1.000	0.990	0.990	0.980
	10	0	0.005	0.004	0.004	0.005	0.006	0.004	0.005
		3	0.040	0.790	0.730	1.000	1.000	1.000	1.000
		5	0.060	0.830	0.790	1.000	1.000	1.000	1.000
20%	2	0	0.006	0.006	0.006	0.007	0.005	0.005	0.005
		3	0.030	0.520	0.480	0.700	0.670	0.660	0.620
		5	0.080	0.590	0.570	0.980	0.970	0.960	0.930
	6	0	0.006	0.007	0.006	0.007	0.005	0.007	0.006
		3	0.030	0.550	0.540	0.990	0.970	0.970	0.960
		5	0.050	0.650	0.630	1.000	1.000	1.000	1.000
	10	0	0.003	0.005	0.003	0.005	0.004	0.004	0.004
		3	0.040	0.750	0.740	1.000	1.000	1.000	1.000
		5	0.060	0.860	0.850	1.000	1.000	1.000	1.000

T^2_{RMVE} based chart was the best. Conversely as the sample size increased T^2_{RMCD} slightly outperformed the T^2_{RMVE}.

The T^2 did not work well, even for small sample size. In order to clarify the performance of the re-weighted robust control chart for different breakdown points of $\gamma = 0.5$ and 0.75 we classify the result based on the sample sizes.

Small sample size: Re-weighted robust control chart with $\gamma = 0.75$ worked better than $\gamma = 0.5$ in high dimensions.

Large sample size: For the small proportion of outlier, the Re-weighted robust control chart for both amount of $\gamma = 0.5$ and 0.75 worked almost similarly. However, if the percentage of outlier in data set was increased the Re-weighted robust control chart with $\gamma = 0.5$ outperformed the other methods. It is worthwhile noting that, when the phase I sample contained higher proportion of outliers, higher value of breakdown point was preferred. This is the main reason for the failure of robust control charts with $\gamma = 0.75$ when there are large numbers of outliers in phase I with large δ^2.

CONCLUSIONS

Standard control charts are widely used in industry to detect the special causes of variation. The common out-of-control status is the occurrence of several outliers in the process. It is well known that the usual parameters estimations are sensitive to the presence of outliers, so the T^2 chart based on these estimators performs poorly. Our advice in this study is replacing the classical mean vector and covariance matrix of the data in the Hotelling's T^2 statistic by the Re-weighted MCD and Re-weighted MVE estimators. These estimators have many advantages, like affine equivariance and better efficiency than the ordinary MCD and MVE estimators used in Vargas (2003), Hardin and Rock (2004) and Jensen *et al.* (2007).

We also recommend generating the T^2_{RMCD} and T^2_{RMVE} control limit for combination of different sample size and dimension via Monte Carlo simulation. Our simulation studies showed that when the process was in-control and the sample size was small, the best estimator was the standard T^2, as noted in the literatures. Nonetheless for large sample size the T^2_{RMCD} and the T^2_{RMVE} performed similar to the classical T^2, chart. On the other hand, when there was outlier in phase I, T^2_{RMCD} and T^2_{RMVE} were more effective than the standard T^2 and ordinary MCD and MVE charts.

In summary, we suggest that when phase I sample has outliers, the RMVE chart is suitable for small sample size and the RMCD chart is the best in the case of large sample size. Moreover, it is better to use $\gamma = 0.5$ with a large sample size (at least 10 times greater than p) to ensure better and consistent performance.

ACKNOWLEDGMENTS

We are grateful to Dr. Chenouri at University of Waterloo, Canada for his constructive comments.

REFERENCES

Butler, R.W., P.L. Davies and M. Jhun, 1993. Asymptotics for the minimum covariance determinant estimator. Ann. Stat., 21: 1385-1400.

Chenouri, S., S. Steiner and A.M. Variyath, 2009. A multivariate robust control chart for individual observations. J. Qual. Technol., 41: 259-271.

Croux, C. and G. Haesbroeck, 1999. Influence function and efficiency of the minimum covariance determinant scatter matrix estimator. J. Multivariate Anal., 71: 161-190.

Davies, L., 1992. The asymptotics of Rousseeuw's minimum volume ellipsoid estimator. Ann. Statist., 20: 1828-1843.

Hardin, J. and D.M. Rocke, 2004. Outlier detection in the multiple cluster setting using the minimum covariance determinant estimator. Comput. Stat. Data Anal., 44: 625-638.

Hotelling, H., C. Eisenhart, H. Hastay and W.A. Wallis, 1974. Techniques of Statistical Analysis. McGraw-Hill, New York, pp: 111-184.

Jensen, W.A., J.B. Birch and W.H. Woodall, 2007. High breakdown estimation methods for phase I multivariate control charts. Qual. Reliabil. Eng. Int., 23: 615-629.

Lopuhaa, H.P. and P.J. Rousseeuw, 1991. Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. Ann. Stat., 19: 229-248.

Pison, G., S. van Alest and G. Willems, 2002. Small sample corrections for LTS and MCD. Metrika, 55: 111-123.

Rousseeuw, P. and B. van Zomeren, 1990. Unmasking multivariate outliers and leverage points. J. Am. Stat. Assoc., 85: 633-639.

Rousseeuw, P. and K. Van Driessen, 1999. A fast algorithm for the minimum variance determinant estimator. Technometrics, 41: 212-223.

Rousseeuw, P.J. and A.M. Leory, 1987. Robust Regression and Outlier Detection. 1st Edn., Wiley, New York, USA., ISBN-10: 0471852333, pp: 352.

Rousseeuw, P.J., 1985. Multivariate estimation with high breakdown point. Math. Statist. Appl., 13: 283-297.

Sullivan, J.H. and W.H. Woodall, 1996. A comparison of multivariate control charts for individual observations. J. Qual. Technol., 28: 398-408.

Tracy, N.D., J.C. Young and R.L. Mason, 1992. Multivariate control charts for individual observations. J. Qual. Technol., 24: 88-95.

Van Aelst, S. and P.J. Rousseeuw, 2009. Minimum volume ellipsoid. Wiley Interdisciplinary Rev. Comput. Stat., 1: 71-82.

Vargas, N.J.A., 2003. Robust estimation in multivariate control charts for individual observations. J. Qual. Technol., 35: 367-376.

Willems, G., G. Pison, P.J. Rousseeuw and S. van Alest, 2002. A robust hotelling test. Metrika, 55: 125-138.

Wisnowski, J.W., J.R. Simpson and D.C. Montgomery, 2002. A performance study for multivariate location and shape estimators. Qual. Reliability Eng. Int., 18: 117-129.