The Effect of Median Based Estimators on CUSUM Chart

Ayu Abdul. Rahman, Sharipah Soaad Syed Yahaya, Abdu Mohammed Ali Atta
School of Quantitative Sciences, UUM College Arts & Sciences, 06010 Universiti Utara Malaysia Sintok, Kedah, Malaysia
ayurahman@uum.edu.my

***Abstract*—Cumulative Sum (CUSUM) chart has been used extensively to monitor mean shifts. It is highly sought after by practitioners and researchers in many areas of quality control due to its sensitivity in detecting small to moderate shifts. Normality assumption governs its ability to monitor the process mean. When the assumption is violated, CUSUM chart typically loses its practical use. As normality is hard to achieve in practice, the usual CUSUM chart is often substituted with robust charts. This is to provide more accurate results under slight deviation from normality. Thus, in this paper, we investigate the impact of using robust location estimators, namely, median and Hodges-Lehmann on CUSUM performance. By pairing the location estimators with a robust scale estimator known as median absolute deviation about the median *(MADn)*, a duo median based CUSUM chart is attained. The performances of both charts are studied under normality and contaminated normal distribution and evaluated using the average run length (*ARL*). While demonstrating an average power to detect the out-of-control situations, the in-control performances of both charts remain unaffected in the presence of outliers. This could very well be advantageous when the proposed charts are tested on a real data set in the future. A case in point is when the statistical tool is used to monitor changes in clinical variables for the health care outcomes. By minimising the false positives, a sound judgement can be made for any clinical decision.**

***Index Terms*—Average Run Length (*ARL*); Contaminated-Normal Distribution; CUSUM Control Chart; Median Based Estimators.**

I. Introduction

As one of the memory-type charts, Cumulative Sum (CUSUM) control chart is known for its reliability in monitoring small shifts in the process mean [1]. The performance of the chart is frequently measured regarding its average run length (*ARL*). The *ARL* is used to gauge how responsive is the chart towards special causes if these variations occur in Phase II. However, calculation of the *ARL* is also governed by normality assumption. Thus, when violated, the chart is expected to signal more frequently than its nominal *ARL* would suggest. In general, this translates to unnecessary process adjustment and loss of confidence in any chart as monitoring tools [2]. To overcome the sensitivity of the usual CUSUM chart in a slight deviation from normality, researchers seek for alternatives in robust CUSUM structure. As such, many have opted to substitute $\bar{X}$ in the plotting statistics with robust location estimators. The idea is to keep the false alarm rate in check upon contamination in the process. For instance, the use of trimmed mean in the CUSUM design has been shown to achieve the said goal [3]. Similarly, reliable *ARL* results were claimed to be achieved when Hodges-Lehmann (*HL*) and tri-mean estimators were applied in CUSUM in nthe on-normal environment [4]. Others have also considered robust estimation when the underlying process mean and variance are unknown. More recently, [5] applied the classical estimators to attain the process parameter in Phase I normal condition, whilst proposing several robust location estimators; trimmed mean, *HL*, tri-mean and median, in the plotting statistics of three memory-type charts; CUSUM, Exponentially Weighted Moving Average (EWMA) and Mixed EWMA-CUSUM charts. Though the finding disclosed that no single chart or estimator is competent in all data condition, it is hard to turn a blind eye on how fast the standard chart performance deteriorates as the level of contamination increases.

In practice, normality assumption is hard to achieve nor do we have the value of the process parameters readily available most of the time. Thus, in this paper, we study the estimation of both location and dispersion process parameter; μ_0 and σ_0, respectively, when Phase I data may contain outliers. To accomplish this, we examine the effect of using median based estimators with CUSUM control structure. Two robust location estimators, namely, median and *HL*, are to be used both for plotting statistics and for deriving $\hat{\mu}_0$. We use only one method to derive $\hat{\sigma}_0$, so that the contrast in the CUSUM performance are merely due to the estimation of μ_0. A consistent estimator of σ_0, known as median absolute deviation about the median (*MADn*) is opted in this article.

The outline of the paper is structured as follows. In Section II, we present the standard CUSUM control chart, follows by the description of the median based estimators in Section III. A step by step approach to construct the proposed method is delineated in Section IV. Section V detailed out the simulation outcomes. The final section; Section VI, summarises the conclusion of this study.

II. The Standard CUSUM Control Structure

Page [6] proposed the idea to measure the accumulative sum of deviation of data from the in-control process mean in two different plotting statistics; the upper ($C_{U,i}$) and the lower ($C_{L,i}$) part as in Equation (1).

$$\begin{aligned} C_{U,i} &= \max\left\{0, C_{U,i-1} + \left(Z_{U,i} - k_U\right)\right\} \\ C_{L,i} &= \min\left\{0, C_{L,i-1} + \left(Z_{L,i} + k_L\right)\right\} \end{aligned} \quad (1)$$

where i defines the subgroup number, $C_{U,0}$ *and* $C_{L,0}$ are the initial values; typically set at 0. The standardized statistics ($Z_{U,,i}, Z_{L,i}$) and the reference values (k_U, k_L) are defined as Equation (2) and (3), respectively.

$$Z_{U,i} = Z_{L,i} = \frac{\hat{\theta}_i - \mu_0}{\sigma_0/\sqrt{n}} \quad (2)$$

$$k_U = k_L = \frac{\delta_{opt}}{2} \quad (3)$$

where $\hat{\theta}$ is the location estimator used to monitor the mean shift, μ_0 and σ_0 are the in-control process parameter, *n* is the sample size and δ_{opt} is defined as a standardized shift in the location where a quick detection is required. An out-of-control signal will be given at time *i*, if $C_{U,i} > h$ or $C_{L,i} < -h$, where *h* is the decision limit.

When process parameters are unknown, common practice is to estimate them based on the sample mean and sample standard deviation. However, these two statistics are easily perturbed by extreme values, which consequently may render the practical use of CUSUM chart meaningless. This leads us to the search of alternatives estimators that could help to alleviate the problem.

III. Robust Location and Dispersion Estimators

The proposed methods of this article are established based on three median based estimators, namely median, *HL* and *MADn*. While the first two estimators in the list are used to measure the location, *MADn* is employed to measure the dispersion. The description each estimator is explained as follows.

A. Median

The estimate provided by median separates the lower half of the data to its upper half. Sample median is computed as in Equation (8).

$$med = \begin{cases} \frac{1}{2}\left[X_{\left(\frac{n}{2}-1\right)} + X_{\left(\frac{n}{2}+1\right)}\right], & if\ n\ is\ even \\ X_{\left(\frac{n}{2}+\frac{1}{2}\right)}, & if\ n\ is\ odd \end{cases} \quad (8)$$

where *n* is the sample size. The efficiency of sample median also rivals the sample mean when tails of the distribution become heavier, irrespective of the sample size [7]. On normal data, the efficiency is set at 64% [8].

B. Hodges-Lehmann (HL)

Identified as the median of pairwise averages, this estimator is proposed by Hodges and Lehmann [9] and defined as in Equation (9).

$$HL = median\left(\left(X_a + X_b\right)/_2\ , 1 \le a \le b \le n\right) \quad (9)$$

It is suited for symmetric models that are prone to outliers as its Gaussian efficiency closely tied-up to the sample mean, measured at approximately 96%. Its breakdown point (*BP*) is relatively lower than the median, placed at 29%, while the former is set at 50% [10]. On that account, some would forgo *HL* and capitalise on another alternative with higher finite breakdown point when dire contamination is speculated.

C. Median Absolute Deviation About the Median (MADn)

This dispersion estimator has made a mark in modern, robust statistical methods due to its 50% *BP* as well as its bounded influence function, with the sharpest possible bound among scale estimators [8]. These merits occasionally outweigh a duo setback experienced by *MADn*; 37% efficiency at Gaussian data and less suitable for asymmetric distributions. It is defined as Equation (10).

$$MADn = \left(1/_{0.6745}\right) med_i|x_i - med_j x_j| \quad (10)$$

The constant $1/_{0.6745}$ in the formula is needed to make the estimator consistent for the parameter of interest, i.e. σ.

The following section describes the design procedure to attain new robust CUSUM charts using the median, *HL* and *MADn*.

IV. The Development of the Proposed CUSUM Control Structure

Rather than assuming data is free from contamination, we recognise the possibility of the presence of outliers. Thus, we propose a duo median based chart for monitoring the location shift. The first chart is the *med*-CUSUM chart; constructed using *med* and *MADn*, while the second chart is identified as *HL*-CUSUM chart; developed using *HL* and *MADn*.

The proposed charts are constructed in two stages; namely Phase 1 and Phase II. The steps to design an optimal *med*-CUSUM chart to detect a shift in location are as follows:

Phase I:

Step 1 Decide on the sample size, *n* and subgroup size, *m*. In this article, we set *m* = 50 and $n \in \{7, 10\}$.

Step 2 Generate the data based on selected distributions (explained later in the next section).

Step 3 Compute the $\overline{Med}$ using the average of the *Med* estimators as in Equation (8).

Step 4 Compute the $\overline{MAD}_n$ using the average of the *MADn* estimators as in Equation (10).

Note that, for the *HL*-CUSUM chart, we compute $\overline{HL}$ using the average of the *HL* estimators as in Equation (9). This refers to the third step in Phase I.

Phase II:

A. To derive optimal parameters

Step 1 Fixed *n*.

Step 2 Fixed in-control *ARL* when data ~*N(0,1)*.

Step 3 Set δ_{opt}

Step 4 Set *k* using Equation (3)

Step 5 Simulate *h* for CUSUM such that the factor would produce value of in- control *ARL* in Step 2.

B. To compute the ARL

Step 1 Generate 15,000 new observations of size *n* from the selected distributions (which is the same as Phase I data distribution). Assume data are in-control.

Step 2 Calculate the *Med* using Equation (8) for each subgroup. This will be the value of $\hat{\theta}_i$ in the plotting statistics.

Step 3 Compute the plotting statistics using Equation (1) and record whether they are within decision limits or not. The respective sample number (when either $C_{U,i} > h$ or $C_{L,i} < -h$) is noted as the in-control run length.

Step 4 Repeat the process for 10,000 simulation runs and compute the *ARL* over the total runs.

Step 5 Introduce shift, $\delta \in \{0.25, 0.5, 0.75, 1, 1.25, 1.5, 1.75, 2.0, 2.5, 3.0\}$.
Step 6 Repeat steps 1 - 4 to get the out-of-control *ARL*.

Note that, for the *HL*-CUSUM chart, we compute $\hat{\theta}_i$ using Equation (9). This refers to the second step in Phase II- *B*.

Once the steps are completed, the simulation outcomes are readily available for discussion.

V. Simulation Outcomes

This section discusses the performance of proposed methods. Two scenarios are considered in this study. First is the ideal condition, where data are assumed to be *N(0,1)*. Second is the non-normal environment, wherein data come from a contaminated normal (CN) distribution. In CN distribution, 95% of the observations come from *N(0,1)*. The remaining 5% of the data are still normally distributed with mean zero, but with a larger σ to illustrate the presence of outliers. In this study, we set $\sigma = 9$.

The results of our simulation study are summarised in Table 2 and 3. To accomplish that, we set the nominal in-control *ARL* = 370 and $\delta_{opt} = 1$. Throughout this article, ARL_0 and ARL_1 are used to denote the in-control and out-of-control *ARL*, respectively. The associated factor (h) for each chart can be referred in Table 1. Along with that, we include the simulation outcome for the standard CUSUM chart as a basis for comparison. The standard chart is constructed using the average of sample mean and average of sample standard deviation in Phase I, while the sample mean is employed in the charting statistics for monitoring process mean in Phase II.

Table 1
Factor (h) of CUSUM charts under $N(0,1)$ at ARL_0 = 370

n	Standard CUSUM chart	*Med*-CUSUM chart	*HL*-CUSUM chart
7	5.048	10.41	12.45
10	5.072	9.22	10.21

The tabulated value in Table 2 gives ARL_0 at a nominal level for all charts since each chart is specifically designed for quick detection of shift size 1 at ARL_0 = 370. Under normality and when $\delta > 0$, the standard chart is highly sensitive to signal the out-of-control situation. Note also that the ARL_1 decreases as *n* increases. Between the two robust charts, *HL* outperforms the median chart irrespective of the sample size.

Table 3 gives the result for all three charts when data are non-normal. The strength of the proposed methods is evident when there are contaminations in the data. In general, the use of robust estimators dictates the in-control performance of CUSUM chart. Both robust charts performances are stable even when outliers are a presence, while the performance of the standard chart is declining markedly for $\delta = 0$. The relationship between *n* and ARL_0 bears little importance here.

Although outlying values in the data have little to no impact on the ARL_0 of the *Med*-CUSUM chart, the chart is highly resistant to the presence of outliers. As such, it is outperformed by the standard chart when $\delta > 0$. Trails behind the standard chart is the *HL*- CUSUM chart.

Table 2
ARL Value for Proposed Charts with with m = 50 at ARL_0 = 370 with $\delta_{opt} = 1$ under *N(0,1)*

n	δ	Standard CUSUM chart	*Med*-CUSUM chart	*HL*-CUSUM chart
7	0	370.248	370.098	370.137
	0.25	26.475	55.711	31.891
	0.5	7.106	18.839	9.643
	0.75	4.090	11.233	5.727
	1	2.967	7.980	4.145
	1.25	2.359	6.266	3.275
	1.5	2.042	5.173	2.731
	1.75	1.831	4.426	2.367
	2	1.592	3.879	2.119
	2.5	1.158	3.367	1.870
	3	1.008	2.884	1.546
10	0	370.102	370.888	369.687
	0.25	17.633	51.703	21.487
	0.5	5.509	19.324	7.231
	0.75	3.360	11.781	4.432
	1	2.472	8.531	3.258
	1.25	2.048	6.702	2.605
	1.5	1.796	5.544	2.204
	1.75	1.513	4.736	3.367
	2	1.237	4.163	2.884
	2.5	1.012	1.414	3.367
	3	1.000	1.067	2.884

Table 3
ARL value for proposed charts with m=50 at ARL_0=370 with $\delta_{opt} = 1$ under *CN* distribution

n	δ	Standard CUSUM chart	*Med*-CUSUM chart	*HL*-CUSUM chart
7	0	220.142	365.707	350.21
	0.25	32.738	67.893	39.293
	0.5	8.221	21.602	11.143
	0.75	4.657	12.568	6.496
	1	3.316	8.945	4.629
	1.25	2.620	6.999	3.643
	1.5	2.205	5.740	3.018
	1.75	1.961	4.886	2.607
	2	1.776	4.255	2.297
	2.5	1.336	3.450	1.982
	3	1.059	2.917	1.763
10	0	231.452	368.21	356.523
	0.25	22.337	64.222	26.677
	0.5	6.339	22.082	8.235
	0.75	3.773	13.393	4.930
	1	2.754	9.543	3.615
	1.25	2.218	7.449	2.881
	1.5	1.936	6.147	2.407
	1.75	1.708	5.249	2.120
	2	1.442	4.598	1.963
	2.5	1.066	3.696	1.656
	3	1.002	3.145	1.224

VI. Conclusion

In this paper, we have studied the performance of median based CUSUM charts. More specifically, we have investigated the effect of a duo median based location estimators; median and *HL*, on the CUSUM structure. To examine the in-control robustness of the CUSUM chart, the process is calibrated using contaminant data of the same type as are subsequently monitored. The goal is to keep the *ARL* relatively close to the nominal level even when outliers are a presence. By pairing each of the proposed location estimators with a high breakdown point scale estimator known as

MADn, this target can be met, remarkably well by the proposed methods.

In general, the proposed charts are expected to be applicable in the manufacturing industry, where the quality of a product is in frequent need of constant monitoring. A more recent trend is to apply control chart in monitoring health care outcomes as the statistical tool is known to be relatively inexpensive, yet powerful enough for overseeing chronic disease. Thus, it vital to minimise false positives and or/ false negatives that could lead to the erroneous clinical decision. This could very well be accomplished via our proposed methods.

ACKNOWLEDGEMENT

The authors would like to acknowledge the work that has led to this paper, which is fully funded the Fundamental Research Grant Scheme of the Ministry of Higher Education, Malaysia at Universiti Utara Malaysia.

REFERENCES

[1] D. C. Montgomery, *Introduction to Statistical Quality Control*, New York: Wiley, USA, 2009, pp: 400-410.

[2] Y. S. Chang and D. S. Bai, "A Multivariate T2 Control Chart for Skewed Populations Using Weighted Standard Deviations," *Quality and Reliability Engineering International*, vol. 21, no. 1, pp. 31-46, 2004.

[3] D. M. Rocke, "Robust Control Charts," *Technometrics*, vol. 31, no. 2, pp. 173–184, May 1989.

[4] H. Z. Nazir, M. Riaz, R. J. M. M. Does and N. Abbas, "Robust CUSUM control charting," *Quality Engineering*, vol. 25, no. 3, pp. 37–41, Jun. 2013.

[5] H. Z. Nazir, N. Abbas, M. Riaz and, R. J. M. M. Does, "A comparative study of memory-type control charts under normal and contaminated normal environments," *Quality and Reliability Engineering International*, vol. 32, no. 4, pp. 1347–1356, 2016.

[6] E. S. Page, "Continuous inspection schemes," *Biometrika*, vol. 41, no. 1, pp. 100–115, Jun. 1954.

[7] F. Figueiredo, and M. 1. Gomes, "The total median in statistical quality control," *Applied Stochastic Models in Business and Industry*, vol 20, no. 4, pp. 339–353. May 2004.

[8] P. J. Rousseeuw and C. Croux, "Alternatives to the Median Absolute Deviation," *Journal of the American Statistical Association*, vol. 88, no. 421, pp. 1273, Jan. 1993.

[9] J. L. Hodges and E. L. Lehmann, "Estimates of Location Based on Rank Tests," *The Annals of Mathematical Statistics*, vol. 34, no. 2. pp. 598–611, Jun. 1963.

[10] M. O. Abu-Shawiesh and M. B. Abdullah, "New Robust Statistical Process Control Chart for Location," *Quality Engineering*, vol. 12, no. 2, pp. 149–159. 1999.