Comparison of Hotelling, MVE and WD for Detecting Outlier in Robust Multivariate Control Chart

Adi Pranata, Kusman Sadik, Erfiani

Abstract— Control chart is the most valuable tool in statistical process control and it is used to monitor the changes in a process. It presents a graphical display of process stability or instability over time. This enable us to instantly understand, whether the process is under control or not. Hotelling T^2 chart is one of the most popular control chart for monitoring multivariate data. In the computation, the parameter of the Hotelling T^2 can be heavily influenced by outliers. In order to decrease the impact of outlier, MVE or WD chart can be conducted as a robust multivariate control chart. In this paper, focused on selecting the most sensitive chart in detecting outliers by comparing signal probability of each method. The results showed that MVE chart has the highest signal probability in uncorrelated data and WD chart is robust in correlated data.

Index Terms— Multivariate control chart, Outliers, Robust, Signal probability.

---

1 Introduction

Quality control is a method used to monitor the product quality and detect problems happened during production process. In controlling quality, one product is expected to have identical characteristics with another. In fact, there are differences among products, called varieties. If the quality control is well-applied, the varieties among products can be detected and fixed soon to produce homogeneous products through controlled production process.

Tool used in controlling quality is control charts. Based on the data used, control charts are differentiated into two which are attribute control chart and variable control chart. In variable control chart, the data are from controlled production process since the characteristics can be measured and analyzed. While based on the characteristics, control charts are differentiated into univariate control chart and multivariate control chart**.** Univariate control chart is a control chart only considering one characteristic. Multivariate control chart is a control chart used to control production process with some characteristics, which either correlated or not, and it can detect changes happened in a production process soon [3]. The strengths of multivariate control chart are efficient, applicable, and understandable. One of common used multivariate control chart is T^2 Hotelling chart. If we compare T^2 Hotelling chart with other types such as MCUSUM or MEWMA, T^2 Hotelling chart has simple and applicable calculating procedure. However, it also has weaknesses.

According to Chenouri *et al* [2], T^2 Hotelling chart is not sensitive with outlier. According to Abu-Shawiesh *et al* [1], outlier can affect significantly in parameter estimator and create uncontrollable production process. Sullivan & Woodall [5] and Vargas [6] make changes to the chart so that robust towards the outliers. Vargas [6] recommends to construct T^2 control chart by using *minimum volume ellipsoid* (MVE) method. MVE method is an effective method to detect outlier. However, MVE is not easily applied because the computation for searching the estimator is difficult and the estimator calculating procedure is not guaranteed from the right data.

To solve that problem, weighted mean vector and mean square successive differences (WD) introduced by Pan & Chen [3] is used in this study. The method is more reliable than MVE method because it is more effective to detect outlier, either random or not by using computation which is easier to use by weighting population mean estimator and modifying the variance-covariance matrix.

In multivariate production process, characteristics among observation could be correlated and uncorrelated. In each situation, outliers are frequently found, in the same time, either with certain pattern or randomly. To find the appropriate method to solve some outlier cases, it is needed to compare one method with another. This study compares MVE method with WD method applied in each correlated and uncorrelated data, considering that both methods are applied for outlier data. Then, both methods are compared with T^2 Hotelling method to find the appropriate method for outliers. The data simulation is used for constructing the data needed in this study.

2 Hotelling Control Chart

First introduced by Harold Hotelling. Basic concept of control chart is monitor production process on several characteristics, both of which correlated or not. The formula of T^2 Hotelling control chart is established from generalized t test.

Let $\boldsymbol{x}_i = (x_{i1}, \dots, x_{ik})'\ j = 1, \dots, m\,; k = 1, \dots, p$ be independently and identically distributed as $N_p\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, thus Hotelling T^2 for single observation (n=1) can be written as

$$\boldsymbol{T}^2 = (\boldsymbol{x}_i - \bar{\boldsymbol{x}})'\, \boldsymbol{S}^{-1} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}) \qquad (1)$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{m}\sum_{j=1}^{m} \boldsymbol{x}_j \qquad (2)$$

$$\boldsymbol{S} = \frac{1}{(m-1)}\sum_{j=1}^{m} (\boldsymbol{x}_j - \bar{\boldsymbol{x}})(\boldsymbol{x}_j - \bar{\boldsymbol{x}})' \qquad (3)$$

3 Minimum Volume Ellipsoid Control Chart

Vargas [6] recommended using the MVE estimator in $\boldsymbol{T}^2$ control chart for detecting any outliers. The basic concept is finding minimum volume ellipsoid which can reach at least half of m points of observation. MVE estimator can be written as

$$T^2_{MVE} = (\boldsymbol{x} - \bar{\boldsymbol{x}}_{MVE})' \boldsymbol{S}^{-1}_{MVE} (\boldsymbol{x} - \bar{\boldsymbol{x}}_{MVE}) \tag{4}$$

The MVE estimators of location and dispersion can be denoted as

$$\bar{\boldsymbol{x}}_{MVE} = (\sum_{j=1}^{h} w_j \boldsymbol{x}_j) / \sum_{j=1}^{m} w_j \tag{5}$$

$$\boldsymbol{S}_{MVE} = \frac{1}{\sum_{j=1}^{m} w_j - 1} \sum_{j=1}^{h} w_j (\boldsymbol{x} - \bar{\boldsymbol{x}}_{MVE})'(\boldsymbol{x} - \bar{\boldsymbol{x}}_{MVE}) \tag{6}$$

where w_j the weight determined 0 or 1.

4 Weighted Mean Vector and Mean Square Successive Differences (WD) Control Chart

Pan & Chen [3] introduced WD method for simplify the MVE computation so that can be efficient applied. WD estimator can be written as

$$T^2_{WD,i} = \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_W\right)' \boldsymbol{S}^{-1}_{WD} \left(\boldsymbol{x}_i - \bar{\boldsymbol{x}}_W\right) \tag{7}$$

$$\bar{\boldsymbol{X}}_{WD} = \frac{1}{\sum_{j=1}^{m} W_j} \sum_{j=1}^{m} W_j \boldsymbol{x}_j \tag{8}$$

$$S_{WD} = \frac{1}{2\sum_{j=2}^{m} W_j} \sum_{j=2}^{m} W_j (\boldsymbol{x}_j - \boldsymbol{x}_{j-1})(\boldsymbol{x}_j - \boldsymbol{x}_{j-1})' \tag{9}$$

5 Comparisons of Detection Performance among Various Control Chart

A control chart is said to be sensitive to outliers if capable of detecting outliers but the parameters have not changed. The sensitivity can be measured by signal probability. Signal probability is probability to detect the signal of out-of-control process. Thus, the various control chart can be compared by signal probability which is written as

$$\frac{\sum Y_i}{n_r} \quad i = 1,2, \dots n_r \tag{10}$$

where the indicator function Y_i equals $T^2 >$ control limit and calculated from n_r equals to number of observation

6 Simulation Result

Data with 50 observations (m) in normally distributed with location parameter $\boldsymbol{\mu}_n$=[0, 0, ..0] and parameter scale $\boldsymbol{\Sigma}_a$ for uncorellated data and $\boldsymbol{\Sigma}_b$ for correlated data which shows that observed characteristics are not correlated. Before constructing control chart, there are some analysis conducted, which are:

6.1 Detecting Outliers

The outlier proportion which is on trial are 5% and 10%. Detecting outliers aims to test whether the generated data contains outliers or not. It is done by using Mahalanobis distance, considering the assumptions:

H_0= A point, not an outlier
H_1=A point, an outlier

From the detection, it shows that the generated data in each characteristic contains outliers.

6.2 Multivariate Normality Test

The data containing outliers are assumed to be distributed normal, so they need to have normality test. The multivariate normality test uses HZ test. Assumptions considered in this study are:

H_0=Data in multivariate normal distribution
H_1=Data not in multivariate normal distribution

Data is considered to be multivariate normal if p> 0.05. The result shows that the data containing outliers both in 5 % and 10% proportion are data in multivariate normal distribution

6.3 Establishment of Control Limit

The control limit from each methods are obtained from the simulation done 100 times. Due to the invariance of the $T^2_{Hotteline}$, T^2_{MVE} dan T^2_{WD} statistics, estimators which are *invariance* or do not change either from transformation or pattern changes, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ in an in-control multivariate normal distribution are assumed to be equal to a zero vector **0** and the identity matrix **I** respectively. The simulation runs for the establishing control limits are shown in the table below

TABLE 1
The Simulated control limit result

Type of Data	Number of characteristics (p)	m=50		
		Hotelling (T^2)	MVE (T^2_{MVE})	WD (T^2_{WD})
Uncorrelated	3	7.47	13.47	7.73
	5	10.45	19.95	11.28
Correlater	3	7.48	13.01	7.82
	5	10.23	19.27	10.94

The simulated control chart above is used to measure the control chart sensitivity. The simulated control chart is established from 0 and 0.95 correlation among characteristics. The control limit determination is based on 95% percentile. It is assuming that control chart has 95% density.

Table 1 shows control limi for correlated data is less than uncorrelated data but it is not to significant because in constructing, the control chart has accommodate the relationship among characteristics. The number of characteristics affect control limit. The more observed cha-recteristics, control limit has increase.

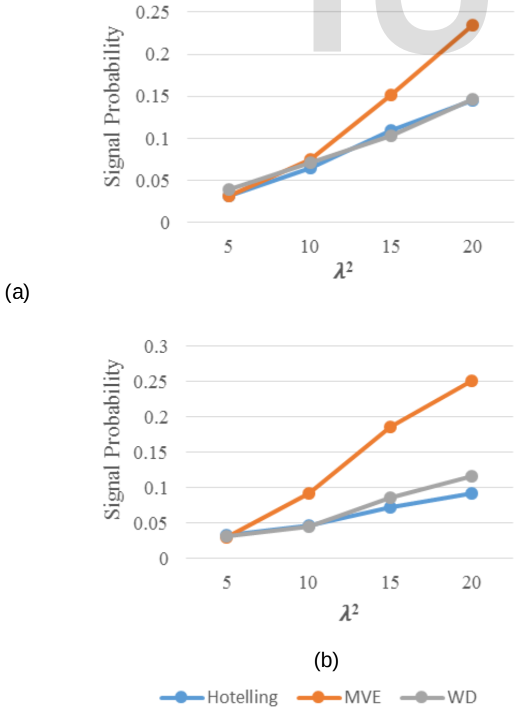
6.4 Comparisons of Detection Performance among Various Control Charts

In establishing robust multivariate control chart, it shows that there are many points which exceed the control limit which then they are used as the signal probability measurement. The Signal probability is calculated in each data condition as simulation plan. The result of simulated *signal probability* calculation and without correlation among characteristics is shown in Table 2.

TABLE 2
THE SIGNAL PROBABILITY WD, MVE AND HOTELLING FOR UNCORRELATED DATA

Number of characteristics (p)	Non-centrality Parameter	Outlier Proportion					
		5%			10%		
		Hotelling	MVE	WD	Hotelling	MVE	WD
3	5	0.032	0.032	0.039	0.032	0.029	0.031
	10	0.065	0.075	0.071	0.047	0.092	0.045
	15	0.109	0.152	0.103	0.072	0.186	0.086
	20	0.145	0.235	0.147	0.092	0.251	0.116
5	5	0.036	0.043	0.041	0.028	0.027	0.031
	10	0.058	0.059	0.058	0.041	0.067	0.044
	15	0.097	0.133	0.087	0.072	0.139	0.065
	20	0.127	0.183	0.125	0.089	0.211	0.089

Signal probability from Table 2 are the result from generating data with $\mu = 0$ and identity matrix Σ_a. The range of the signal probability is come from 0.031 to 0.251. It is means that the signal probability has low value. According to Table 2, along with increasing of outlier proportion, the signal probability has decreased in same characteristics. On the other words, the control chart's sensitivity is depends on outlier proportion. Signal probability for Hotelling, MVE and WD for any situation displayed in graph.



(a)

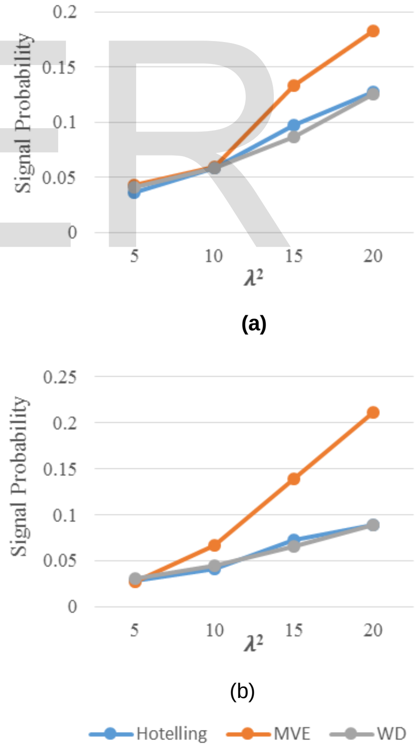
(b)

Figure 1. Signal probability for uncorrelated data with p=3 for (a) outlier proportion = 5% and (b) outlier proportion = 10%.

In Figure 1(a), signal probability of Hotelling, MVE and WD chart increase along with the changes of λ^2value. It shows that Hotelling, MVE and WD method are able to detect outliers if p=3 and outlier proportion is 5%. MVE chart is the most sensitive chart because it has the highest signal probability among others. Moreover, WD chart is the second lead but the difference was not significant so that the sensitivity chart WD does not differ much with Hotelling chart in detecting outliers. Same idea in Figure 1(b), MVE is the most sensitive chart to detect outliers which has highet signal probability in 10% outlier proportion.

The influence of the proportion of outlier sensitivity control chart visible when the value of signal probability of each method in the proportion of 5% compared with the proportion of 10%. On MVE chart, in line with outlier proportion increase, the signal probability increases. Inverse, it is not applicable in WD and Hotelling chart. Their signal probability move downward when the porportion increases.

The Signal probability is also considered in p=5 to check the number of characteristics with some changes towards the sensitivity of multivariate robust control chart. The signal probability chart for p=5 with the move of parameter as far as λ^2.



(a)

(b)

Figure 2. Signal probability for uncorrelated data with p=5 for (a) outlier proportion = 5% and (b) outlier proportion = 10%..

In Figure 3, along with the changes of λ^2, the signal probability of WD and MVE chart is moving upward. Compared to other charts, MVE has highest signal probability with 5 characteristics. Different from other methods, the *signal probability* of Hotelling chart is low and the *signal probability* value reduces when λ^2=20 in 10% outliers. It shows that Hotelling chart is not able to detect outliers well.

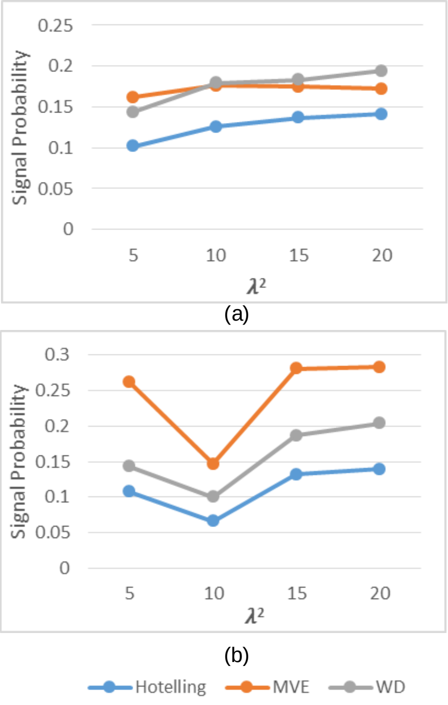
The sensitivity of robust control chart is also affected by number of observed characteristics, therefore the *signal probability* of each multivariate *robust* method towards the characteristics are 3 and 5 at the same outlier proportion.

Measurement of signal probability of correlated data shows in Table 3.

TABLE 3
The Signal Probability WD, MVE and Hotelling for Correlated Data

Number of characteristic (p)	Non-centrality Parameter	Outlier Proportion					
		5%			10%		
		Hotelling	MVE	WD	Hotelling	MVE	WD
3	5	0.102	0.162	0.144	0.107	0.262	0.143
	10	0.126	0.176	0.179	0.066	0.147	0.100
	15	0.137	0.175	0.183	0.132	0.281	0.187
	20	0.141	0.172	0.194	0.139	0.283	0.204
5	5	0.139	0.217	0.203	0.149	0.330	0.199
	10	0.157	0.234	0.245	0.159	0.363	0.207
	15	0.164	0.237	0.266	0.162	0.383	0.211
	20	0.174	0.235	0.277	0.167	0.381	0.215

Tabel 3 shows the signal probability for correlated data range form 0.102 to 0.383. Comparing with Table 2, signal probability in Table 3 higher than on Table 2. In other words, the three control charts are more sensitive to detect outliers in correlated data than when they applied in uncorrelated data. Different from Table 2, Table 3 shows that the signal probability keeps moving upward along number of characteristics' escalation. Comparison along three methods is displayed in graph below:



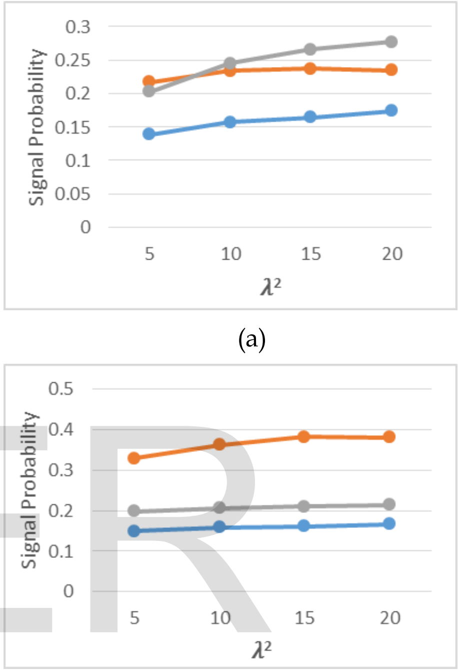
(a)

(b)

Figure 3. Signal probability for correlated data with p=3 for (a) outlier proportion = 5% and (b) outlier proportion = 10%.

Figure 3(a) shows, the signal probability of WD chart keeps moving upward along the changes of λ^2 in proportion 5% meanwhile MVE do not show any sensitivity to outliers with signal probability value which move slightly downward. WD chart is more sensitive than MVE and Hotelling because it has bigger signal probability value.

Figure 3(b) shows that Hotelling, MVE and WD could not detect the outliers properly because their signal probability reduces in $\lambda^2 = 10$. For the comparison of correlated data's signal probability for p=5 is shown in Figure 4.



(a)

(b)

Figure 4. Signal probability for correlated data with p=5 for (a) outlier proportion = 5% and (b) outlier proportion = 10%.

Figure 4 (a) for p=5, WD chart is keeps moving upward when the non-centrality parameter changes and has biggest score in signal probability but MVE proved to insensitive because in $\lambda^2 = 15$ has lower signal probability than $\lambda^2 = 10$. Despite of MVE has the highest signal probability shown in Figure 4 (b), MVE chart is not applicable to detect outliers because weaken in $\lambda^2 = 20$. Based on Figure 3 and 4, WD chart is most suitable method to apply in correlated data.

Comparing Table 2 and 3, signal probability of Hotelling, MVE and WD in correlated data are higher than uncorrelated data. One of the reasons is the control limit in correlated data is lower than uncorrelated data. Based on all graphs above, Hotelling chart has the lowest signal probability. It is sign that Hotteling is not a robust control chart when outliers come.

5 Conclusion

Based on simulation, MVE chart is a robust control chart for uncorrelated data and WD chart is a robust control chart for correlated data because consistently keeps moving upward when non-centrality parameter increases. Besides that, robustness of a control chart is depends on nuber characteristics in data, outlier proportion and the type of data (correlated or uncorrelated). In uncorrelated data, sensitivity control chart is come down when the number of characteristics increases. Different from uncorrelated data, signal probability of correlated data increases when the number of characteristics increases. The proportion of outliers is one indicator of the ability to detect outliers for control chart. In the data are uncorrelated, the greater the proportion of outliers, the ability to detect outliers getting down. Unlike the data are not correlated, sensitivity control chart data is correlated increases with increasing proportion of outliers.

REFERENCES

[1] Abu-Shawiesh M, George F, Kibria B. 2014. A “Comparison of Some Robust Bivariate Control Charts for Individual Observation”. International Journal of Quality Research 8(2) 183 – 196.

[2] Chenouri S, Variyath A, Steiner S. 2007. “A Multivariate Robust Control Chart for Individual Observation”.

[3] Pan J, Chen S. 2011. “New Robust Estimators for Detecting Non-Random Patterns in Multivariate Control Chart: A Simulation Approach”. Journal of Statistical Computation and Simultation vol 81, No 3. March 2011.

[4] Rousseeuw J, Zomeren B. 1990. “Robust Distances: Simulations and Cutoff Values”. Direction in Robust Statistics and Diagnostics, Part II, edited by W. Stahel and S, Weisberg, Springer-Verlag.

[5] Sullivan J, Woodall W. 2000. “Change-Point Detection of Mean Vector or Covariance Matrix Shifts Using Multivariate Individual Observations”. IIE Transactions, 32 pp. 537–549.

[6] Vargas NJA. 2003. “Robust Estimation in Multivariate Control Charts for Individual Observations”. J. Qual. Technol. 35 (2003), pp. 367–376.