ISSN 1823-626X

Journal of Fundamental Sciences

available online at http://jfs.ibnusina.utm.my

Robust start up stage for beltline moulding process variability monitoring using vector variance

Rohayu Mohd Salleh* and Maman A. Djauhari

Department of Mathematics, Faculty of Science, Universiti Teknologi Malaysia, Skudai, Johor.

Received 11 March 2010, Revised 4 May 2010, Accepted 13 May 2010, Available online 25 May 2010

ABSTRACT

One of the primary problems encountered in monitoring the variability of beltline moulding process in an automotive industry is the estimation of parameters in the start-up stage. This problem becomes more interesting because the process is in multivariate setting and must be monitored based on individual observations, i.e., the sample size of each subgroup is 1. This paper deals with a robust estimation of location and scale during the start-up stage. For this purpose, we use Mahalanobis distance in data ordering process. But, in data concentration process, we use vector variance (VV). This method is highly robust and computationally efficient. Its advantage in monitoring the variability of beltline moulding process will be compared with the non-robust method.

| breakdown point | covariance determinant | Mahalanobis distance | robust estimation | vector variance |

http://dx.doi.org/10.11113/mjfas.v6n1.179

1. INTRODUCTION

It is known that a successful of monitoring process in Phase II depends on a successful analysis during start up stage (SUS) or Phase I (Jensen et. al, 2005). Even though the two phases are both dedicated to identify out-of-control states, each phase has a unique objective. If SUS is used to estimate parameters, Phase II consists of monitoring future observations by using information from in-control historical data set (HDS) in SUS to determine whether or not the process continues to be in stable condition. Consider the situation when random sample data are stored in $n \times p$ matrix where n and p are the number of observations and variables, respectively. Let X_i be the vector representing the i-th row. We assume that X_i; $i = 1, 2, ..., n$ are independent and follow a multivariate normal distribution. These data vectors will be used in start-up stage to obtain an in-control data subset which will be used to estimate the process parameters. Since μ and Σ are unknown, they are replaced with an appropriate estimators mean vectors, $\vec{\bar{X}}$ and covariance matrix, S. These estimators are needed to monitor the process variability right after a future data vector or, equivalently, individual observation is available. Since the data is in multivariate setting, it is not easy to identify the outliers during the start-up stage as the analysis will be done simultaneously.

$$\vec{\bar{X}} = \begin{bmatrix} \bar{X}_1 \\ \bar{X}_2 \\ \vdots \\ \bar{X}_p \end{bmatrix}, \; S = \begin{bmatrix} s_{11} & \cdots & s_{1p} \\ \vdots & \ddots & \vdots \\ s_{p1} & \cdots & s_{pp} \end{bmatrix}$$

Derquenne (1992) stated the measure of identification multivariate outliers is created by a technique of transforms the random vectors to be random variables so that candidates of outlier will be seen more clearly. The most popular transformation is the Mahalanobis squared distance (MSD). A large value of MSD may indicate that corresponding observation is an outlier. Explained by Hadi (1992), outliers do not necessarily have large value of MSD and not all observations with large MSD value are necessarily outliers. These problems are known as masking and swamping effect due to the fact that mean vector and covariance matrix are not robust.

To handle this problem, the method of robust estimator is applicable as theoretical foundations of the construction of distance which is robust MSD. This paper is organized as follows. Section 2 and 3 presents classical approach and robust approach in SUS. We present an illustrative example on real life data of beltline moulding to demonstrate the effectiveness of robust approach compared with classical approach.

Corresponding author at: Department of Mathematics, Faculty of Science Universiti Teknologi Malaysia,81310, Skudai, Johor, Malaysia.
E-mail addresses: rohayusalleh@yahoo.com (Rohayu Bt Mohd Salleh)

2. CLASSICAL APPROACH

Classical method based on MSD is powerful when there is only one out-of control point. Its power will decrease if more than one out-of control points are present in the data (Hadi, 1992). It is sensitive not only to the shift in mean vector but also in covariance matrix (Tracy, 1992). Any shift in mean vector and or covariance matrix will lead to unstable process.

2.1 Classical distance and distribution

The classical distance is generated from the arithmetic mean. The arithmetic mean is an estimation of the classical mean which is computed from the whole sample. Let $X_1, X_2, ..., X_n$ be a random sample from *p*-variate distribution where the second moment exists. The sample mean vector and sample covariance matrix are, respectively,

$$\vec{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n} \vec{X}_i \text{ and } S = \frac{1}{n-1}\sum_{i=1}^{n} (\vec{X}_i - \vec{\bar{X}})\left(\vec{X}_i - \vec{\bar{X}}\right)'.$$

The MSD is a classical distance which measures each observation $\vec{X}_i$ to $\vec{\bar{X}}$ given by sample covariance S, and formulated by

$$d_i^2\left(\vec{X}_i, \vec{\bar{X}}\right) = \left(\vec{X}_i - \vec{\bar{X}}\right)' S^{-1}\left(\vec{X}_i - \vec{\bar{X}}\right) \quad \text{for all } i = 1, 2, ..., n$$

Based on MSD, the in control data is determined by plotting the value in control chart. Presence of one or more extreme data or called outliers changes the arithmetic mean $\vec{\bar{X}}$ significantly, and the distance increases. It is sensitive not only to the shift in mean vector but also in covariance matrix (Tracy, 1992). Any shift in mean vector and or covariance matrix will lead to unstable process. Since our aim in SUS is to check whether the observations are fall within the control limit, MSD is distributed as Beta distribution. Gnanadesikan and Kettenring (1972) based on results of Wilks (1962). Specifically,

$$d_i^2 \sim \frac{(n-1)^2}{n}\beta\left(\frac{p}{2}, \frac{n-p-1}{2}\right)$$

Knowing the distribution of MSD, it is possible to construct the control limits. It is given by

$$LCL = \frac{(n-1)^2}{n}\beta\left(1-\frac{\alpha}{2}, \frac{p}{2}, \frac{n-p-1}{2}\right)$$

and

$$UCL = \frac{(n-1)^2}{n}\beta\left(\frac{\alpha}{2}, \frac{p}{2}, \frac{n-p-1}{2}\right).$$

3. ROBUST APPROACH

Classical estimation methods will not yield appropriate control limits if there are unusual data points in SUS. Robust estimation methods have advantage over classical methods in that they are not unduly influenced by outlier data points. Jensen et al. (2005) discussed the use of robust MSD method based on minimum volume estimate (MVE) and minimum covariance determinant (MCD) criteria in start up stage. Both criteria were introduced by Rousseeuw (1985) have good properties, which are affine equivariant and have a high breakdown point if the data set is in general position. Later on, in order to improve its computational efficiency, Rousseeuw (1999) introduced a faster algorithm called FMCD. However, see for example Werner (2003) and Djauhari (2007) this algorithm is still cumbersome when the data set is of high dimension. Djauhari (2007) introduced a new robust estimator method called as minimum vector variance (MVV). This estimator still has the same structure like FMCD but use the different concept. Like FMCD, in the first step, we still use robust MSD as data ordering. In data concentration step, instead of calculate generalize variance; we change the procedure of data concentration by using vector variance (VV). The objective of FMCD is to find the best subset of h that having the minimum covariance determinant or generalized variance (GV). This objective will be the stopping rule of FMCD. However, the objective and stopping rule of MVV is to find the best subset of *h* that having the minimum vector variance (VV). VV is the sum of square of all elements of the covariance matrix. As a measure of multivariate variability, VV performs much better than GV (Djauhari, 2008) for small shift in covariance matrix. There are two advantages of using VV as multivariate data concentration. First, the computation is far more efficient even for large matrix size than GV. Second, VV does not need the condition of non-singularity of covariance matrix. Unlike VV, the GV needs the condition that the covariance matrix must be non-singular.

3.1 Data concentration using VV

Consider a random vector data set of p-variate normal observations which is in general position.

- ❖ Compute the mean vector $\vec{\bar{X}}_{H_{old}} = \frac{1}{h}\sum_{i=1}^{h} \vec{X}_i$ and covariance matrix $S_{H_{old}} = \frac{1}{h}\sum_{i=1}^{h} (\vec{X}_i - \vec{\bar{X}})\left(\vec{X}_i - \vec{\bar{X}}\right)'$. These estimators are calculated based on the data subset of $h = \frac{n+p+1}{2}$. Define the relative distances

$d^2_{H_{old}}(i) = \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right)' S^{-1} \left(\vec{X}_i - \bar{\vec{X}}_{H_{old}}\right)$ for all $i = 1, 2, ..., n.$

- Sort these squared distances in increasing order, $d^2_{H_{old}}(\pi(1)) \le d^2_{H_{old}}(\pi(2)) \le \ldots \le d^2_{H_{old}}(\pi(n))$ where π is a permutation on $\{1, 2, ..., n\}$.

- Define $H_{new} = \{\vec{X}_{\pi(1)}, \vec{X}_{\pi(2)}, \ldots, \vec{X}_{\pi(h)}\}$. Compute $\bar{\vec{X}}_{H_{new}}, S_{H_{new}}$ and $d^2_{H_{new}}$.

- If $Tr(S^2_{H_{new}}) = 0,$ repeat the above procedure. If $Tr(S^2_{H_{new}}) = Tr(S^2_{H_{old}}),$ the process is stopped. Otherwise, the process is continued until the *k*th iteration.

- Let us denote the location and covariance matrix given by MVV as follows;

$$\vec{T}_{MVV} = \frac{1}{h}\sum_{i\in H}\vec{X}_i$$

and

$$S_{MVV} = \frac{1}{h}\sum_{i\in H}\left(\vec{X}_i - \vec{T}_{MVV}\right)\left(\vec{X}_i - \vec{T}_{MVV}\right)'$$

Based on MVV, robust MSD is then defined by

$$d^2_{MVV}\left(\vec{X}_i, \vec{T}_{MVV}\right) = \left(\vec{X}_i - \vec{T}_{MVV}\right)' S^{-1}_{MVV}\left(\vec{X}_i - \vec{T}_{MVV}\right)$$

for all $i = 1, 2, ..., n$

The MVV estimator, see Herwindiati et al. (2007), are calculated based on the best subset of $h = \frac{n+p+1}{2}$ data points, having minimum $Tr\left(S^2\right)$among all possible subsets in order to get the best estimates of mean vector, $\bar{\vec{X}}$ and covariance matrix, S . The data concentration of MVV is accurate as well as FMCD.

3.2 Robust distance and distribution

By using robust estimates, it gives MSD with unknown distributional properties. However, using robust estimates gives MSD with unknown distributional properties. In Hardin and Rocke (2005), an approximate result for MSD based on location and scatter are derived. The distribution of S_{MVV} can be approximated by

$$mc^{-1}S_{MVV} \sim Wishart_p(m, \Sigma)$$

where m and c are the unknown parameter. Therefore, Hardin and Rocke (2005) approximate the distribution of extreme distances by approximating the distribution of the MVV shape by Wishart, so that we can apply the *F* distribution.

$$d_i^2 \sim \frac{pm}{c(m-p+1)}F_{p,m-p+1}$$

The control limits can be expressed as

$$LCL = F_{\left(1-\frac{\alpha}{2}, p, m-p+1\right)}\left(\frac{c(m-p+1)}{pm}\right)$$

and

$$UCL = F_{\left(\frac{\alpha}{2}, p, m-p+1\right)}\left(\frac{c(m-p+1)}{pm}\right)$$

where m and c are the unknown parameter. The parameter c can be estimated by $c = \frac{P\left(\chi^2_{p+2} < \chi^2_{p,h/n}\right)}{h/n}$, p is the number of variables and $\hat{m} = \frac{2}{\widehat{CV}^2}$. $\widehat{CV}$ is the estimated coefficient of variation of the diagonal elements of the MVV scatter estimator.

4. ILLUSTRATIVE EXAMPLE

One of the existing problems in any automotive industry is in the production process of beltline moulding. Beltline moulding over the outer lip of the drip rail prevents water from leaking into the car. If the lip is quite short, beltline moulding often will not position well. On the other hand, if the lip is longer the window glass will not move smoothly (Bon, 2008). This type of problem is not easily solved by applying standard procedure of manufacturing since the variability among the materials, machine processes, ambient conditions and end products exist and cannot be avoided.

The beltline moulding data are stored in $n \times p$ data matrix where n and p are the number of observations and variables, respectively. In this paper, we use the data in Bon (2008) which consist of $n = 57$ and $p = 8$. Since the beltline moulding data set is multivariate setting, it is not easy to realize outliers in SUS.

Firstly, we will show the difference in SUS between classical estimation and robust estimation. The mean vector and covariance matrix based on classical estimation are

$$\bar{\bar{X}} = \begin{pmatrix} -0.0526 \\ 0.5491 \\ -0.0439 \\ 0.1228 \\ 0.1193 \\ 0.2395 \\ 0.5561 \\ 0.3518 \end{pmatrix}, S = \begin{pmatrix} 0.05977 & & & & & & & \\ -0.01304 & 0.07513 & & & & & & \\ 0.11970 & -0.00196 & 0.41563 & & & & & \\ 0.01636 & 0.06864 & 0.13392 & 0.53554 & & & & \\ 0.04130 & 0.00939 & 0.11367 & -0.02147 & 0.09051 & & & \\ 0.01466 & 0.01347 & 0.13301 & 0.09766 & -0.00738 & 0.40078 & & \\ 0.00537 & -0.00357 & 0.00273 & -0.01617 & 0.00836 & -0.02484 & 0.04402 & \\ -0.00388 & -0.01638 & -0.02314 & -0.03701 & -0.00117 & -0.00695 & 0.00320 & 0.02482 \end{pmatrix}.$$

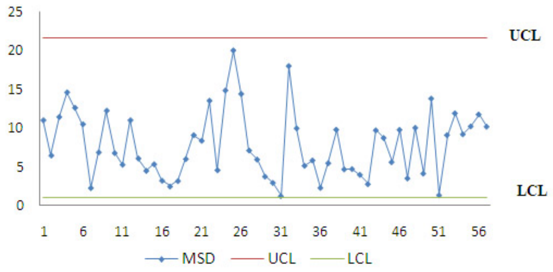

Figure 1 SUS based on classical MSD

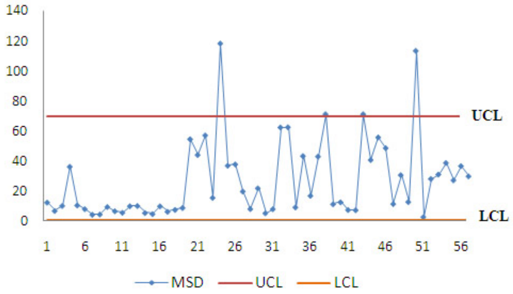

Figure 2 SUS based on robust MSD

The determinant value is $|S| = 3.266\times10^{-9}$.

Figure 1 visualizes the SUS based on classical approach or non-robust MSD. From the table of Beta distribution with degree of freedom *p* = 8 and probability of false alarm 0.0027, UCL = 21.6134 and LCL = 0.9959. We see that no observation lies outside the control limits. However, some observations are of large variation. In the next sub-section we continue to further analyse the start-up stage using robust MSD to see whether masking and swamping effects occur. The mean vector and covariance matrix based on robust estimation by using MVV data concentration

$$\bar{\bar{X}} = \begin{pmatrix} 0.0375 \\ 0.6226 \\ 0.1775 \\ 0.4200 \\ 0.1825 \\ 0.2500 \\ 0.5600 \\ 0.2550 \end{pmatrix}, S = \begin{pmatrix} 0.05391 & & & & & & & \\ -0.00049 & 0.02118 & & & & & & \\ 0.11602 & -0.01776 & 0.44881 & & & & & \\ 0.02671 & 0.01111 & 0.08640 & 0.39853 & & & & \\ 0.01003 & -0.00064 & 0.03182 & -0.06345 & 0.03691 & & & \\ 0.05158 & -0.01000 & 0.20895 & -0.06184 & 0.02724 & 0.26026 & & \\ -0.00303 & 0.00713 & -0.01634 & 0.00176 & 0.00124 & 0.00329 & 0.04016 & \\ 0.002303 & -0.00117 & 0.01960 & -0.01024 & 0.00009 & 0.00250 & 0.01021 & 0.01103 \end{pmatrix}.$$

The determinant value of robust estimation is 2.74×10^{-11} . The value of determinant calculated from robust estimation is small compared to classical estimation. The variability of covariance matrix of robust estimation is less. The above parameter estimation, is calculated from sample subset, $h = 33$.

Figure 2 shows the SUS control chart based on robust MSD. From the table of *F* distribution, with p =8, probability of false alarm = 0.0027, *c* = 1.06455 and m_{pred} = 41.67, then UCL = 51.8580 and LCL = 1.3322. Observations 24, 38, 43 and 50 have the largest MSD and lie outside the control limits. Furthermore, observations 1 to 19 have small values of MSD but observations 20 to 51 have large variation. This figure presents the information that cannot be provided by non robust MSD chart. Figure 1 does not signal any out of control condition. Consequently, we already obtained the best parameter estimation by using robust approach.

Removing all four outlying observations and recalculating the parameter estimates with $n = 53$ and $p = 8$, the new mean vector and sample covariance were obtained by using the classical estimation of equation

(1). The corresponding control limits for this sample of size 53 follow the approximate distribution of

$$\frac{(n-1)^2}{n} d^2\left(\vec{X}_i, \bar{\bar{X}}\right) \sim Beta\left(\frac{p}{2}, \frac{(n-p-1)}{2}\right)$$

as explained by Hardin and Rocke (2005). Then, with probability of false alarm = 0.0027, the UCL = 21.4812 and LCL = 0.9985.

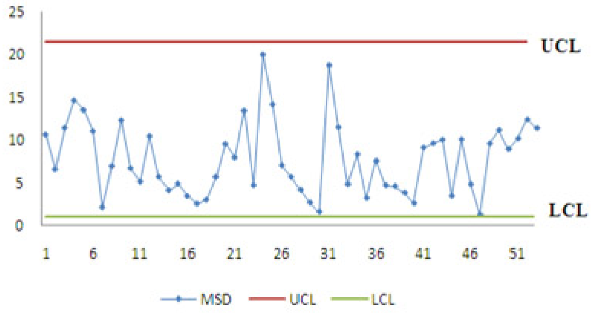

Figure 3 Robust control chart with modified data set

Reconstructing the control chart, we observe as in Figure 3 that none of the observations are outside the control limits. The new control chart has been established by eliminating the special cause of variation from outlying observations at Figure 2.

This beltline moulding data illustrates the effectiveness of the robust MVV estimator compared to the classical estimator in detecting process variability.

5. CONCLUSION

The monitoring variability during SUS by using beltline data indicates that VV is effective as CD as one of robust method. It is practically approved to be used as one of method for engineering consideration in control the production process because of the computational efficiency and easy to implement. Since the engineering experiments are quite particular with number of sample, VV can be applied in both conditions, either sub grouped observations or individual observations.

6. PROBLEM TO SOLVE

In view of the fact that in this paper, we just show the approach of MVV estimator during Phase I or SUS, we will further analyse in the process Phase II monitoring.

ACKNOWLEDGMENT

The authors thank to Universiti Teknologi Malaysia (UTM) under the grant vote 77536, Ministry of Higher Education of Malaysia and Universiti Tun Hussein Onn Malaysia (UTHM) for support of this research. .

REFERENCES

[1] A.S. Hadi (1992). Journal of Royal Statistical Society, Vol. B, 53, 761 - 771.

[2] A.T Bon (2008). Process quality improvement on beltline moulding manufacturing. PhD Thesis. Universite De La Rochelle, France.

[3] C. Derquenne (1992). *J.Siam*, 34(2), 323-326.

[4] D.E. Herwindiati, M.A. Djauhari, and M. Mahsuri (2007). *Journal of Communication in Statistics - Computation and Simulation*, 36, 1287-1294.

[5] J. Hardin, and D.M. Rocke (2005). *Journal of Computational and Graphical Statistics*, Vol 14, No 4, 928– 946

[6] M.A. Djauhari (2007). *Journal of Applied Probability and Statistics*, 2, 139-155.

[7] M.A. Djauhari, M. Mashuri, D. E. Herwindiati (2008). *Journal of Communication and Statistics – Theory and Methods*, 37, 1742-1754.

[8] M. Werner (2003). Identification of multivariate outliers in large data sets. Phd Thesis, University of Colorado at Denver.

[9] N.D. Tracy, and J. C. Young (1992). *Journal of Quality Technology*, Vol 24, No 2, 88-95.

[10] P. J. Rousseeuw (1985). Multivariate estimation with high breakdown point. In: Grossman, B. W., Pflug, G., Vincze, I., Wertz, W., eds. *Mathematical Statistics and Applications*. D. Reidel Publishing Company, pp. 283–297.

[11] R. Gnanadesikan and J.R. Kettenring, J.R. (1972). *Biometrics* 28, 81-124.

[12] W.A. Jensen, J.B. Birch, and W.H. Woodall (2005). High breakdown point estimation methods for phase I multivariate control charts. Department of Statistics, Virginia Polytechnic Institute and State University Blacksburg