The Control Chart of Data Depth Based on Influence Function of Variance Vector

Suwanda Idris¹⁺, Lisnur Wachidah², Teti Sofiyayanti³, Erwin Harahap⁴

¹,²,³Departement of Statistics, Faculty of Math and Science Universitas Islam Bandung
⁴Departement of Mathematics, Faculty of Math and Science Universitas Islam Bandung, Jalan Ranggamalela No. 1 Bandung, West Java Indonesia.

*suwanda@unisba.ac.id, †idris100358@gmail.com

Abstract. The concept of data depth has been used for multivariate control charts. One of the advantages of data depth is that statistics constructed independent of the distribution of quality characteristics or is known as free distributions. One of them is depth based on Mahalanobis distance. Individual Mahalanobis distance is essentially formed by the ratio of generalized variance (GV) calculated by not involving the individual to be calculated distances to the mean by GV calculated from the complete data. The Control charts constructed with Mahalanobis Depth can be visualized (exploration) is plot depth over depth, so it is known DD-diagram and create control limits as control chart Shewhart. The main disadvantage of Mahalobis distance lies in GV, the value of GV depends on the sample size and the number of quality characteristics, in addition two different covariance matrices can give the same GV value. Departing from the weakness of GV, this research will be focused on the formation of vector data depth control diagrams (VV), because VV can resolve the GV weaknesses. Based on the simulation study, it shows that both DD-GV and DD-VV control charts give the same pattern in ongoing process of in-control that is forming a straight line through point (0,0) and (1,1), for correlation between weak variables, but for correlation between variables are very strong, the DD-GV control does not expressly state the in control process. The application process control of data in both diagram medical product, indicate that there is a symptom of mean vector or covariance matrix.

1. Introduction
A control chart is a standard tool used to monitor the quality of a process or identify instability in a manufacturing process. In practice, the quality of a product is determined by the interaction of multiple correlated characteristics, this is a multivariate natural phenomenon. Therefore appropriate techniques must be used in monitoring the quality of multivariate processes [4].

The first multivariate Shewhart control chart was introduced in 1947. Its formation was based on statistics $T^2$ - Hotelling. Then a number of multivariate control chart are designed for different situations such as the MCUSUM and MEWMA diagrams. These diagrams classic for monitoring have been built under a number of certain assumptions [1].

The performance of multivariate control charts in particular individual control chart is very dependent on the hypothesis that the distribution underlying the quality of the process is multivariate normal. However, in practice this hypothesis might not be fulfilled. Alternative procedures are needed to overcome this limitation. Based on the idea of data depth, [6] proposed various control chart similar to individual control charts, average, and cusum. Then [4] proposes a visual procedure called depth over depth (DD) diagram that uses a data depth plot to monitor the quality of multivariate data and does not require assumptions about the distribution underlying the process. This graphical method provides visualization of changes in location and / or on a scale between an empirical sample of a reference sample. Furthermore, these diagrams are easier to interpret and quick adjustments to the quality of multivariate processes through the use of control limits are adjusted to detect each out-of-control signal.
The type of depth used by [4,6] is the Depth distance of the Mahalanobis. The Mahalanobis distance is derived from the calculated generalize variance ratio without involving the individual \( i \)th \( (GV_i) \) on the generalize variance sample calculated from the complete data. \( GV \) is calculated from the determinant of the sample covariance matrix. The disadvantage of \( GV \) is that the value is close to zero if there is a large correlation between variables or the sample size is relatively small compared to the number of variables, and two different covariance matrices are possible to have the same \( GV \) [7].

To overcome drawbacks of \( GV \), its can use vector variance [9,10]. Vector variance is calculated from the trace square of the covariance matrix. The depth built is through the function of the influence of \( VV \) samples. Furthermore, the control chart obtained is called a control chart data depth variance vector. This control chart is made as proposed by [4], namely through a procedure visualization of depth over depth. In this paper, the definition of data depth, \( GV \) depth data, \( VV \) depth data and control chart depth over depth are presented. The final section presents an illustration with simulation data and real data.

2. The Methodology

2.1. Data Depth

Assume that \( p > 1 \) the characteristics of each product are used to determine the quality of the product. This process is considered in control if the measurement follows the specified quality distribution (which is needed by the customer or the design engineer). For example, \( G \) denotes a specified \( p \)-dimensional distribution, and for example \( Y_1, Y_2, \ldots, Y_m \) is a random observation of \( G \) with size \( m \). This sample, generally referred to as reference samples in the context of quality control, and that are considered as measurements of products produced by the in-control process. Let, \( X_1, X_2, \ldots \) are a new observations (empirical sample) of the manufacturing process. Assume that following distribution \( F \). Based on observations \( X_i \), it will be determined whether the product quality is deteriorating or whether the process is out of control. This means that \( X_i \) does not meet \( G (.\) which is specified in a certain sense. Therefore, we need to compare \( F \) with \( G \). The statistics used is to describe certain aspects of the difference between \( G \) and \( F \) are based on the idea of the depth of the data, hence we begin by explaining some of the concepts of data depth.

One type of depth is based on the distance of the Mahalanobis. Next, how deep is the point \( y \) with respect to distribution \( G \), how small is the square distance of the mean vector

\[
MD_G (y) = \frac{1}{1 + (y - \mu)^\top \Sigma^{-1} (y - \mu)}
\]

where \( \mu \) and \( \Sigma \) each is the mean vector and covariance matrix of \( G \). The empirical version of (1) is

\[
MD_G (y) = \frac{1}{1 + (y - \bar{y})^\top S^{-1} (y - \bar{y})}
\]

where \( \bar{y} \) and \( S \) each of them is the mean vector and covariance matrix of the sample \( Y_1, Y_2, \ldots, Y_m \). It can be shown that it fulfills the property affine invariant [6]. Furthermore depth is written in notation \( D_G (.) \). Assume that \( G \) and \( F \) are continuous distributed absolutely.

Obviously, the data depth induces an outer sequence from the sample center point if the depth values for all points are calculated and compared. More specifically, if we arrange all \( D_G (y_i) \) in ascending order and use \( Y_{(1)} \) to show the sample point that corresponds to the smallest depth value \( f^{th} \), then \( Y_{(1)}, Y_{(2)}, \ldots, Y_{(m)} \) are the order statistic of \( Y_i \), by being the most central point. The smaller the order (or rank) of the point, the more remote the point is regarding the distribution underlying \( G (.) \).
Depth data (1) and (2) are called $GV$ depth data or $DDGV$. It is clear that $DDGV$ requires that the covariance matrix or nonsingular ones. Singularity of the covariance matrix can occur if there is a perfect linear or almost perfect correlation between variables, and or sample size (n) is smaller than the number of variables (p). Alternatively, you can use vector variance data, $DDVV$.

The dept data Vector variance is derived from the influence function of vector variance. It has been shown by [11] that the effect of an observation vector on vector variance measured by:

$$FVV = -\left(y - \mu \right)^\top \Sigma \left(y - \mu \right)$$  \hspace{1cm} (3)

Form (3) its value is always exist, even though the covariance matrix is singular. The population depth-$GV$ (1) or sample (2) can be used for $VV$ depth by replacing the Mahalanobis distance with the formulation distance as shown in (3). Furthermore, the data depth calculated from the function of the influence of $GV$ is denoted by $DDGV$, and calculated from the effect of $VV$ is denoted by $DDVV$.

It should be noted that several other control chart derived from VV have been discussed in [2], [3], [8], [11], and [12].

2.2 Plot Depth with Depth (DD).

Let $X = \{X_1, X_2, \ldots, X_n\}$ and $Y = \{Y_1, Y_2, \ldots, Y_m\}$ are two random samples, each from $F$ and and $G$, where $F$ and $G$ are two continuous distributions in $\mathbb{R}^p$. Comparing two samples can be studied in the framework of testing the null hypothesis,

$$H_0 : F = G$$

The alternative hypothesis is very dependent on the problem being investigated, whether composite or simple. Visually to explore whether the data of the two samples support or reject the null hypothesis can use plot depth with depth, referred to as DD plots. In particular, DD plots are plots from $DD\left(F_n, G_m\right)$, where

$$DD\left(F_n, G_m\right) = \left\{(DF_n(x), DG_m(x)) , x \in \{X \cup Y\}\right\}$$ \hspace{1cm} (4)

Note that $DD\left(F,G\right)$ and $DD\left(F_n, G_m\right)$ are always subsets of $\mathbb{R}^2$ no matter how big the dimension $p$ is from the data. Two-dimensional graphs of DD plots are easy to visualize and they are easy to use tools for graphical comparisons of multivariate samples.

If $F = G$, then $DF(x) = DG(x)$ for all $x \in \mathbb{R}^p$, and thus the $DD\left(F,G\right)$ is a line segment only on the line 45° in the DD plot, from (0, 0) to $(\max, DF(t), \min, DG(t))$.

Associated with controlling the process, for example in the case of in-control processes in the sense that there is no shift in parameters from both location vectors and dispersion matrices control chart $DD$ will show a straight-line pattern through points $(0,0)$ and $(1,1)$.

3. The Control Chart Use Plot Depth with Depth (DD)

The procedures for making control chart $DD$ are as follows [5]:

a) Determine historical data of size $n$ taken in the previous process in-control. Furthermore, this data is called reference data.

b) In the ongoing process, take data with size $m$. Next is called empirical data.

c) Calculate the mean vector and covariance matrix using reference data.

d) Calculate the mean vector and covariance matrix using empirical data.
e) Combine reference data and empirical data.
f) Calculate Depth from the combined data using the mean vector and covariance matrix in (c).
g) Calculate Depth from the combined data using the mean vector and covariance matrix in (d).
h) Plot the results of (f) data depth of the opponent with data depth results (g).
i) If the Mahalanobis distance is used, the plot obtained we call the DDGV control chart, if we use the function of the vector variance effect we call the DDVV control chart.
j) If the plot results show data distribution that is not through point (0,0) and (1,1), it can be concluded that it can be assumed that there has been a change in distribution or process in an out of control state.
k) We draw the conclusions objectively point (j) through hypothesis testing on regression coefficients between Depth reference data ($y$) with empirical depth data ($x$), $y = \alpha + \beta x + \varepsilon$ with the formulation of hypotheses $H_{01}: \alpha = 0$ and $H_{02}: \beta = 1$ for Depth samples.

4. The Data Depth Control Chart for Simulation Data

The data generate is carried out for known process parameters as well as unknown process parameters with each created in four scenarios.

4.1. For Parameters of Process Known

The first scenario, the simulation is done by generating reference data ($X$) from $N_10(0, I)$, with the size $n = 400$ and empirical data ($Y$) also from $N_10(0, I)$, with the size of $m = 300$. By using the depth (2.3) pair, the results show that the control Chart DDGV (figure 1 (a)) and control chart- DDVV (figure 1 (b)) shows scatter forming a straight line and through point (0,0) and (1,1), in other words the process of control and describing diagrams is indeed the actual situation like that.

![DDGV plot](image)

**Figure 1.** a) DDGV plot and b) DDVV plot with reference data and empirical data derived from the same multivariate normal distribution, $X \sim N_{10}(0, I), Y \sim N_{10}(0, I), n = 400, m = 300$.

The second scenario is the same as the first scenario with the first variable averaging shifting to 1. Both charts shows that the process is out of control because both the DDGV control chart (figure 2 (a)) and DDVV control chart (figure 2 (b)) does not show the scatters that forms a straight line and through point (0,0) and (1,1), and indeed the actual situation like that.
Figure 2. DDVV plot with reference data $X \sim N_{10}(0, I)$ and empirical data with the first mean variable shifting to 1, $n = 400$, $m = 300$.

The third scenario is the same as the first scenario, only here are the reference and empirical sample sizes $n = 15$ and $m = 8 < p = 10$, respectively. Both chart shows that the process is in control because both the DDGV control chart (figure 3 (a)) and the DDVV control chart (figure 3 (b)) show the points drawn in a straight line and through a point $(0,0)$ and $(1,1)$, and indeed the actual circumstances like that. Thus the sample size has no effect when the distribution parameters are known.

Figure 3. (a) DDGV plot and (b) DDVV plot for Third Scenario Data

The fourth scenario, is the same as the first scenario, only here there is a change in correlation between the first variable and the second variable to 0.9. Both chart shows that the process is out of control because both the DDGV control chart (figure 4 (a)) and the DDVV control chart (figure 4
(b) do not show points that form a straight line and through a point (0, 0) and (1,1), and indeed such actual conditions, namely the covariance matrix changes (multivariate dispersion shifts).

Figure 4. (a) DDGV Plot (b) DDVV Plot for Fourth Scenario Data

4.2. Parameters of Process Unknown

The fifth scenario, the sample depth for the first scenario with sample size \( n = 50 \) and \( m = 40 \). The results show that the DD-GV control diagram (figure 5(a)) process is out of control, whereas according to DDVV (figure 5 (b)) shows in control processes. The actual situation is an in control process, and this is not tracked by the DDGV control chart.

Figure 5. (a) DD GV sample Plot (b) DD VV Sample Plot for Data Scenario 1.

Sixth scenario, sample depth for the first scenario with the mean process the first variable shifts to 2 and the size of the inferential sample and empirical sample are \( n = 40 \), \( m = 30 \), respectively. The results show that DDGV control chart (figure 6 (a)) and DDVV chart (figure 6 (b)) both show scatter not forming a straight line through points (0,0) and (1,1), or processes out of control.
Figure 6. (a) DDGV plot, (b) DDVV plot for 6th Scenario Data

Seventh scenario, third sample depth scenario with sample size \( n = m = 8 \). The results show that the DDGV control chart (figure 7 (a)) depth value tends to be zero, because \( n = m = 8 < p = 10 \) and DDVV (figure 7 (b)) the value varies but does not go to zero because it is not affected by sample size. From the results of the partial test, the DDGV control chart shows the out of control process, which is actually an in-control process, while the DDVV control diagram shows the in control process, which is actually the in control process. Thus the DDVV control diagram can still work even though the sample size is smaller than the number of variables.

Figure 7. (a) DD GV plot, (b) DDV for data Scenario 7

The eighth scenario, the seventh scenario with the first variable shifts to 1. The partial test of regression coefficient, results show the DDVV chart explicitly the out of control at the 5%.
Ninth scenario, sample depth from fourth scenario (coefficients of variables 1 and 2 are made at 0.99). The results show that the DDGV control chart (figure 9 (a)) does not state that the process is in control, while the DDVV control chart (figure 9 (b)) states the in control process as indicated by scattering not forming a straight line through point (0,0) and (1,1). In this case the DDVV control diagram expressly states the process in control and indeed the actual state of the process in control.

5. The Application on Real Data
Data is obtained from a leading manufacturer and supplier of products made from absorbent cotton gauze and non woven material. This company operates in the medical field that wants to ensure that all products that are used for medical human beings, will be safe and effective in accordance with European and International standard, see [4]. The variable used is the strength of the sealing edge in the left, right, up and down direction. Reference and empirical data are 40 in size each, observations are presented in table 1.

Estimator of average vector and covariance matrix:
From reference data.
\[ \bar{x}_r = \begin{pmatrix} 3.8625 \\ 3.64725 \\ 3.6605 \\ 3.655 \end{pmatrix} \text{ and} \]
From Empirical data.

\[
\begin{pmatrix}
0.2699 & -0.0051 & 0.0944 & -0.0316 \\
-0.0051 & 0.5328 & 0.1168 & 0.0120 \\
0.0944 & 0.1168 & 0.5582 & 0.2094 \\
-0.0316 & 0.0120 & 0.2094 & 0.4950
\end{pmatrix}, \quad \hat{GV}_F = 0.0284, \hat{VV}_F = 1.8558
\]

The results are based on DDGV control chart and DDVV control chart showing the out of control process, see figure 10 (a) (DDGV chart) and figure 10 (b) (DDVV Chart), because both scatter deviating from (0,0) and (1,1). There has been a shift in location parameters and or multivariate dispersion.

![Figure 10. (a) DDGV plot and (b) DDVV plot for Seal Strength Data](image)

6. Conclusion and Recommendation

The DDVV control chart can be used as an alternative in controlling multivariate processes. The DDVV control chart expressly signals the in-control process when the correlation between variables is very large and the sample size is small, for the actual process also in control. The implementation of the seal strength control shows the out of control process according to both the DDGV and DDVV control chart at a real level of 5%.

In this study, it still needs to be continued to determine the theoretical properties of DDVV, for example, determine the limit distribution of DDVV and its convergence speed. The control limits of DDVV must also be determined so that the Average Run Length can be calculated.

Acknowledgements

The authors are deeply thankful to editor and reviewers for their valuable suggestions to improve the quality of this manuscript. This article was funded by Lembaga Penelitian dan Pengabdian kepada
Masyarakat (LPPM) at Universitas Islam Bandung, Indonesia. The author, Suwanda, therefore, acknowledge with thanks LPPM technical and financial support.

References

[1] JA Alloway 1995 Visual evaluation of multivariate control chart assumptions in Proceedings of the American Statistical Association, pp. 49–54, The Section on Quality and Productivity, American Statistical Association

[2] Dwidaman and Suwanda 2011 Diagram Kontrol VV yang direduksi Thesis Program Magister Statistika Terapan Universitas Padjadjaran Bandung

[3] E Dinihayati and Suwanda 2011 Diagram Kontrol Dispersi Multivariat untuk Ukuran Sampel Kecil Thesis Program Magister Statistika Terapan Universitas Padjadjaran Bandung

[4] M Hajlaoui 2010 A Graphical Quality Control Procedure Using Data Depth Advances and Applications in Statistics 19 2 p97

[5] R Liu and K Singh 1993 A Quality Index Based on Data Depth and Multivariate Rank Tests Journal of the American Statistical Association 88 p252

[6] RY Liu 1995 Control Charts for Multivariate Processes Journal of the American Statistical Association 90 432

[7] DC Montgomery 1991 Introduction to Statistical Quality Control 2nd Edition. John Wiley & Sons New York

[8] Salam and Suwanda 2010 Penentuan Batas-batas Diagram Kontrol Dispersi Multivariat VV dengan Metode Bootstrap Thesis Program Magister Statistika Terapan Universitas Padjadjaran Bandung.

[9] Suwanda and MA Djauhari 2003 A New Concept In Monitoring Multivariate Process Variability Newsletter Data Analysis Research Group Departement of Mathematics Institut Teknologi Bandung 2 2

[10] Suwanda 2004 Bagang Kendali $Tr\left(S^2\right)$ untuk Mengendalikan Dispersi Multivariat Prosiding Konferensi Nasional Matematika XII Denpasar Indonesia p558

[11] Suwanda 2008 Implementasi Variansi Vektor Pada Pengontrolan Dispersi Multivariat Proses Dengan Ukuran Sub Grup Tunggal Konferensi Nasional Matematika XIV dan Kongres Himpunan Matematika Indonesia Jurusan Matematika FMIPA UNSRI Palembang

[12] Suwanda 2016 A Synthetic Control Chart Reexpression Vector Variance for Process Multivariate Dispersion Journal of Mathematics and Statistical Science 2 2