Comparison on the Performance of Several Outlier Detection Methods in Univariate Circular Wrapped Normal Sample

Nur Syahirah Zulkipli¹, a and Adzhar Rambli¹, b

¹Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA, 40450, Shah Alam, Selangor, Malaysia.

a syahirahzulkipli1994@gmail.com and b adzhar_rambli@tmsk.uitm.edu.my

Abstract. This study focuses on detecting a single outlier in circular data generated from a wrapped normal (WN) distribution. The discordancy tests of M, A and G₁ statistics are used to detect single outlier in simulated data generated from wrapped normal distribution. The purpose of this study is to make a comparison on the performance of these statistics via Monte Carlo simulation by obtaining the proportion of correct outlier detection for each statistic. In this study, Splus-language and R-language programming are used to carry out the simulation study. The power performance of these statistics have been investigated and the result revealed that these statistics performed better as the increment in contamination value, λ and the value of concentration parameter, ρ gets larger and close to 1 in the case of small and large sample size, n. In general, the A statistic is found to be outperformed the M and G₁ statistics in all cases. As an illustration, a practical example is included in this study by using the Kuantan wind direction dataset.

Keywords: Circular; Wrapped normal distribution; Outliers; Discordancy test.

1. Introduction

Directional data is a set of data that contains observations with directions [1]. Directional data can be in two dimensions known as a circular data or three dimensions known as a spherical data where it depends on the number of variables in the study. It is a subdiscipline statistics that deals with sample spaces. Sample space is in circumference of a circle or a sphere and since the data is in that measurement, the standard ‘linear’ method used in analysing univariate and multivariate data cannot be applied [2]. In many scientific studies, directional data commonly exist but unfortunately the techniques and specialized statistical techniques are not widely known and appreciated by others [1]. That is why many experts from various area have faced difficulties in analysing their data which in a circumference of a unit circle. For instance, applied scientist like biologist, geologist, social and behavioural scientist who is working with directional data was facing a difficulty when applying the usual linear methods since it is inappropriate [1]. According to [1], it makes the typical linear techniques produce a misleading or meaningless result due to dissimilar of statistical methods and measures in directional data.

As stated by [3], the detection of outliers can reveal surprising or unpredictable information in many crime areas for example in fraud activities detection, criminal behaviour detection and computer intrusion detection. Outlier can be defined as an observation that deviates far away from the other observations and it will become doubtful observation [4]. In other words, an outlier is an observation that has a large distance from other values in a random sample from a population. The observation that separate far from others may occur due to some reasons like cause by human or measurement error. This observation can influence the inferences process. Therefore, this problem may affect the inferences process and it may have a proper work to be done by trying to detect such outliers.
However, the existence of the outlier in linear data is slightly different with the data in circular unit. The problem of the outlier in directional data is a bit different from linear data. In circular data cases, the distance on how far the observation from the mean in directional data setup should be decided by using circular distance. The value of outlier in circular data is differ from linear data. In a circular case, outliers do not need be too large or too small from others and it could be in the “central” of the dataset [1]. Outlier is common to appear in dataset due to several reasons such as human error, systematic error, environmental phenomenon and others. Consequently, it will give a big impact to the estimation and sometimes it can influence other observation values. In circular data, the problem may arise when there are outliers appear in a dataset that follow circular distribution [5]. According to Kaya [6], in statistics, the issue of outliers frequently arises and the concern on this problem had less attraction among academicians and researchers during the last century. Recently, the issue on the outliers is becoming a hot issue among researchers on identifying and handling the problems related to the outliers.

Up to now, only few literatures are found related to the outlier identification in circular data. This is because, the problems of outlier in circular data have not been fully discovered especially for a data that follow wrapped normal distribution. Furthermore, most of the method of detecting outlier in circular data like discordancy test are focusing on detecting a single outlier. However, there is a study proposed a new outlier detection method $G_a$ statistic that be able to detect not only single outlier but also multiple and patches of outliers in circular sample [13]. There are few studies on comparing outlier detection methods in univariate wrapped normal sample that had been conducted [12]. However, the studies that had been conducted have not include $G_I$ statistic. Until now, there is no study on comparing $G_I$ statistic with the other statistics on detecting single outlier in wrapped normal sample. A wrapped normal (WN) distribution is obtained by wrapping a $N(\mu, \sigma^2)$ distribution around the [1]. Asymmetric unimodal distribution which is obtained by wrapping a normal distribution with mean $\mu$ and variance $\sigma^2$ around the circle is called as WN distribution [7]. Hence, this study interested to compare the performance of several outlier detection methods which are the $M$, $A$ and $G_I$ statistics in circular data that come from the wrapped normal distribution and then applying it into a real dataset.

2. Methodology

2.1. Discordancy Test

There are several discordancy tests introduced by previous researchers. Mardia and Jupp [8] introduced the $M$ statistic for discordancy test in circular data and then followed by Collet [9] who introduced the discordancy test of $C$ and $D$ statistics. The $A$ statistic later proposed by Abuzaid et al. [10] and recently, Mohamed et al. [11] introduced the $G_I$ statistics based on spacing theory in circular data. The statistics $M$, $C$, $D$ and $A$ are the discordance tests that can assessing the single angular outlier. However, the $G_a$ statistics are proposed as the discordance test to detect not only a single outlier, but it can detect a multiple or a patch of outliers in circular data.

The discordancy test can be concluded whether the suspicious observation is an outlier by comparing the test statistics with the cut-off point. The null hypothesis is rejected when the test statistics is larger than the cut off-point. Note that, the cut-off point differs for different circular probability distributions (every distribution has their own cut-off point). The cut-off point is the percentage points under the null hypothesis of no outliers in the circular dataset various values of the concentration parameter of the following distribution.

2.1.1. $M$ Statistic

$M$ statistic is proposed by Mardia [8] and the statistic had been pointed as a statistic of discordancy which is given by $M' = \min_i \left\{ \frac{n-1-R_{(i)}}{n-R_{(i)}} \right\}$ However, the statistic was reformulated by Collet [9] in equation (1)
\[ M = 1 - M' = \max \left\{ \frac{R_{i-1}-R+1}{n-R} \right\} = \frac{R_q - R + 1}{n - R}, \]

where \( R_q = \max \{ R_{(i-1)} \} \), \( R = \sqrt{C^2 + S^2} \), \( C = \sum_{i=1}^{n} \cos \theta_i \) and \( S = \sum_{i=1}^{n} \sin \theta_i \)

If the value of \( M \) statistic exceeds a cut-off point, then the \( i \)th observation corresponding to the maximum value of \( M \) is identified as outlier. The cut-off point for \( M \) statistic for WN sample is available in [12].

2.1.2. A Statistic

The \( A \) statistic is proposed by Abuzaid et al. [10] is based on the summation of the circular distances from a point of interest to another points and the statistic is defined as

\[ A = \max \left\{ \frac{D_{ij}}{2(n-1)} \right\}, j = 1, \ldots, n, \]

where \( D_{ij} = \sum_{i=1}^{n} d_{ij} \) and \( d_{ij} = 1 - \cos(\theta_i - \theta_j) \). Thus, this test of discordance shows that it is used circular distance as a measure of dissimilarity. If the value of \( A \) statistic exceeds a cut-off point, then the \( i \)th observation corresponding to the maximum value of \( A \) is identified as outlier. The cut-off point for \( A \) statistic for WN sample is available in [12].

2.1.3. \( G_1 \) Statistic

The \( G_a \) statistic is the latest discordancy test and it was developed by Rambli [13]. The development of \( G_a \) statistic is very well known for detecting a single, multiple as well as a patch of outliers. This study primarily focused on the case when \( a = 1 \). Thus, the \( G_1 \) statistic here looks for maximally isolated point \( \theta_i \) which is attributed as an outlier in the circular data. Suppose \( \theta_1, \theta_2, \ldots, \theta_n \) are (i.i.d) circular observations from a WN distribution. Thus, the step to obtain \( G_1 \) statistic given as follow.

1) Firstly, order the observations as \( \theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(n)} \).
2) Secondly, for a choice of \( a \)-step spacing, calculate \( G_{ai} \), \( i = 1, 2, \ldots, n \) as given in equation (3)

\[ G_{ai} = \theta_{(i+a)} - \theta_{(i)} \text{ for } i = 1, 2, \ldots, n-a \text{ and} \]
\[ G_{ai} = 2\pi - \theta_{(i)} + \theta_{(i+a) - n} \text{ for } i = (n+1) - (n+2) - a, \ldots, n \]  

(3)

3) Thirdly, define \( G_i = \min \left( G_{ai}, G_{a,1-a} \right) \) for \( i = 1, 2, \ldots, n \), which is the smaller of the \( a \)-step spacing on either side of \( \theta_i \).
4) Lastly, define \( G_a = \max \left( G_i \right), i=1,2,\ldots,n. \)

If the value of \( G_1 \) exceeds a cut-off point, then the \( i \)th observation corresponding to \( \max \left( G_i \right) \), \( i=1,2,\ldots,n \) is identified as outlier. The cut-off points for \( G_1 \) statistic for WN sample is available in [14].

2.2. Performance of the discordancy tests

The performance of the statistics to detect an outlier in circular sample need to be tested and a selected measure by [9] are used to compare the performance of the tests. The discordancy test can be known whether it is good or not by looking on how well it is performed. As stated by [15] and [16], a good test should have (i) a high power functions, (ii) a high probability of identifying a contaminating value as an outlier when it is in fact an extreme value where an extreme value is defined as a point with the maximum circular deviation and (iii) a low probability of wrongly identifying a good observation as discordant. In addition, a good test is expected to have:
1. High P1, where let $P1 = 1 - \beta$ be the power function where $\beta$ is the Type-II error.
2. High P5, where let $P5$ the probability that the contaminant point is defined as discordant given that it is an extreme point.
3. Low P1- P3, where let P3 the probability that the contaminant point is an extreme point and is identified as discordant.

The power performance of the $M$, $A$ and $G_1$ statistics are examine using proportion outlier correct detection ($\text{proportion.simu}$). Thus, the performance of the discordancy tests in this study will be expected to have High P5 where the value of proportion outlier correct detection is approaching to 1. The value of proportion is in the range of $[0,1]$. The proportion outlier correct detection is obtained by

$$\text{proportion.simu} = \frac{\text{success}}{\text{number of simulation}}$$

where “success” is number of dataset that correctly detects the outlier at the right position and exceeds the value of cut-off point. The performance is studied and evaluated by plotting the proportion of outlier correct detection against contamination value, $\lambda$. Then, the performance of $M$, $A$ and $G_1$ statistics will be compared through a graph of proportion outlier correct detection against contamination value, $\lambda$. Samples are generated in such a way that $(n-1)$ of the observations that come from $WN (\mu, \rho)$ and one observation from $WN (\mu + \lambda \pi, \rho)$, where $\lambda$ is the degree of contamination with the range $[0,1]$. The value of $\mu = 0$ is fixed and use different values of $\rho$ in the range $[0.7, 0.975]$.

Besides that, several values of sample sizes also used in the range $[10, 100]$ to study the performance of the $M$, $A$ and $G_1$ statistics with different level of $\rho$ at certain sample sizes. The $M$, $A$ and $G_1$ statistics in each random sample are calculated based on equation (1), (2) and (4) respectively. The process of simulation in obtaining the proportion of outlier correct detection is carried 3000 times using 5% percentile level for these statistics. The statistic is accurately identified the observation as an outlier if it correctly detects the outlier at the right position and exceeds the value of cut-off point. The simulation is repeated 3000 times and the percentage of correct performance is estimated by using equation (5).

3. Results

3.1. Performance of the discordancy tests

This study focusses on detection of a single outlier. The results of Monte Carlo simulation of $M$, $A$ and $G_1$ statistics performances are showed in Figure 1 - Figure 6. The performance of $M$, $A$, and $G_1$ statistics for the case when small sample size, $n=20$ and large sample size, $n=100$ are plotted in Figures 1(a)-1(c) and Figures 2(a)-2(c) respectively. From small to large sample sizes, Figures 1 and Figure 2 clearly show that the performance of the $M$, $A$, and $G_1$ statistics are performing better as the value of $\rho$ increases. By looking for the case when the value of $\rho$ is 0.975, the $M$, $A$, and $G_1$ statistics perform well even at lower contamination levels compared to other values of $\rho$.

Next, the performance of these three statistics are evaluated by looking at different perspective for the case when $\rho=0.9$ and 0.975 at different $n$. Figure 3 shows the performance of $M$, $A$, and $G_1$ statistics for the case $\rho=0.9$ is fixed, meanwhile, plots in Figure 4 are the performance when $\rho=0.975$ is fixed. In Figure 3(a), $M$ statistic perform well when $n$ is large. Somehow, for the case $\rho=0.9$, Figures 3(b) and 3(c) suggest that $A$ and $G_1$ statistics are performed well when $n$ is small. When $\rho$ is increased to 0.975, Figures 4(a)-(c) suggest that the performance of these statistics is parallel with the findings in 3(a)-(c). However, it clearly can be seen in Figures 4(a)-(c) where the performance of these statistics become much better and the differences in the performance of small and large sample size become less noticeable.
Figure 1(a). The performance of $M$ statistic for small sample size $n=20$

Figure 1(b). The performance of $A$ statistic for small sample size $n=20$

Figure 1(c). The performance of $G_1$ statistic for small sample size $n=20$

Figure 2(a). The performance of $M$ statistic for large sample size $n=100$

Figure 2(b). The performance of $A$ statistic for large sample size $n=100$

Figure 2(c). The performance of $G_1$ statistic for large sample size $n=100$
Figure 3(a). The performance of $M$ statistic for $\rho=0.9$

Figure 3(b). The performance of $A$ statistic for $\rho=0.9$

Figure 3(c). The performance of $G_1$ statistic for $\rho=0.9$

Figure 4(a). The performance of $M$ statistic for $\rho=0.975$

Figure 4(b). The performance of $A$ statistic for $\rho=0.975$

Figure 4(c). The performance of $G_1$ statistic for $\rho=0.9$
The performance of $M$, $A$ and $G_1$ statistics are evaluated simultaneously and to see the comparison more clearly, the graphs are plotted in Figures 5(a)-(b) and Figures 6(a)-(b) for different cases. In this study, the comparison on the performance of $M$, $A$ and $G_1$ statistics is made based on small and large $n$. The value of concentration parameter, $\rho = 0.9$ and 0.975 are used in the study to compare the performance of these statistics. Results for the case $n = 20$ and $n = 100$ with a fixed value of $\rho = 0.9$ are plotted in Figures 5(a) and 5(b) respectively. The curves in Figure 5(a) revealed that $M$ statistic has the worst performance among these statistics for the case small $n$. However, as $n$ increases to 100, the performance of $M$ statistic has become no difference with $G_1$ statistic as shown in Figure 4.6(b). The results are found that $A$ statistic outperforms the others for the case small and large $n$ when $\rho = 0.9$.

Next, the value of $\rho$ is increased to 0.975. The results are obtained for small and large $n$. Figures 6(a) and 6(b) show the performance of these three statistics for the case $n = 20$ and $n = 100$ with a fixed value of $\rho = 0.975$ respectively. The graphs in Figures 6(a) and 6(b) show that $A$ statistic is more powerful than $M$ and $G_1$ statistics since the performance of $A$ statistic higher than the others for small and large $n$. The curves in Figures 6(a) and 6(b) found that as the $\rho$ gets larger and close to 1, the performance of $M$ statistic still lower than both $A$ and $G_1$ statistics for the case of small $n$. However, the performance of $M$ statistic gets better and slightly surpass the performance of $G_1$ statistic when $n$ is increased to 100 for the case $\rho = 0.975$ as shown in Figure 6(b).

Therefore, we can conclude that, as $n$ increase and $\rho$ close to 1, the differences in the performance of $M$ and $G_1$ statistics become less noticeable. This indicates that the larger the $n$, the less there is to choose between $M$ and $G_1$ statistics when the $\rho$ is 0.9 and 0.975. In general, $A$ statistic is more powerful to detect single outlier compared to $M$ and $G_1$ statistics in univariate wrapped normal sample.

Figure 5(a). Performance of $M$, $A$ and $G_1$ statistics when small sample size $n=20, \rho=0.9$

Figure 5(b). Performance of $M$, $A$ and $G_1$ statistics when small sample size $n=100, \rho=0.9$

Figure 6(a). Performance of $M$, $A$ and $G_1$ statistics when small sample size $n=20, \rho=0.975$

Figure 6(b). Performance of $M$, $A$ and $G_1$ statistics when small sample size $n=100, \rho=0.975$
As summary, the $M$, $A$ and $G_1$ statistics are performed better as the contamination value, $\lambda$ increases, the value of concentration parameter, $\rho$ gets larger and close to 1 for small and large sample size, $n$. For the case when $\rho = 0.975$ the $M$, $A$, and $G_1$ statistics perform well even at lower $\lambda$ compared to other values of $\rho$. In the case sample size is small, $A$ and $G_1$ statistics are more powerful in detecting single outlier in wrapped normal sample. In contrary, $M$ statistic shows that this statistic become a powerful test when the sample size is large. $A$ statistic outperforms the other statistics and found that the performance of $M$ statistic inferior than the $A$ and $G_1$ statistics. However, as the $n$ and $\rho$ get larger, the $M$ statistic gets better and perform as well as the $G_1$ statistic.

3.2. Real data application
This study considered the Kuantan wind direction data measured in unit radian from the year 1999 to 2008. The ten years Kuantan wind direction data is used, and this data is obtained from the Malaysian Meteorological Service Department which available in study Rambli et al. (2012) [12] as shown in Table 1. The circular plot of the data is plotted as in Figure 7 and it shows the dataset has possible outlier.

| Year | Mean surface wind direction (radian/degree) |
|------|------------------------------------------|
| 1999 | 0.28707 / 16.45º                        |
| 2000 | 1.46071 / 83.69º                        |
| 2001 | 0.87509 / 50.12º                        |
| 2002 | 1.64563 / 94.29º                        |
| 2003 | 1.56786 / 89.83º                        |
| 2004 | 1.33478 / 76.48º                        |
| 2005 | 1.80266 / 103.28º                       |
| 2006 | 2.15736 / 123.61º                       |
| 2007 | 1.73430 / 99.37º                        |
| 2008 | 1.67275 / 95.84º                        |

**Table 1.** Kuantan wind direction data

![Circular plot of Kuantan wind direction](image)
Table 2. Summary of discordancy test using $M$, $A$ and $G_1$ statistics for Kuantan wind direction

| Statistics | Test value | Observation, $\theta$ | Cut-off point, 95% | Detection |
|------------|------------|-----------------------|-------------------|-----------|
| $M$        | 0.59       | 1                     | 0.6386            | Not an outlier |
| $A$        | 0.37       | 1                     | 0.6365            | Not an outlier |
| $G_1$      | 0.59       | 1                     | 0.7198            | Not an outlier |

Figure 7 shows that one observation in quadrant one is become a candidate of outlier since it deviates far away from the others. The observation that become the candidate of outlier is the Kuantan wind direction for the year 1999. This suspicious observation is at 1st position and the value of this observation is 0.28707 radian or 16.45º. Table 2 provide the results for detecting the outlier and it can be concluded that the suspicious observation is an outlier if the test value is greater than the corresponding cut-off points at 5% level of significance. It is found that the observation at 1st position cannot be declared as an outlier since the test value for $M$, $A$ and $G_1$ statistics are less than their cut-off points. Moreover, the value of contamination, $\lambda$ for this dataset approximately near to 0.2 which is too small. Therefore, it has low possibility to indicates that the candidate of outlier is an outlier significantly.

4. Conclusions

This study has contributed to circular data analysis especially wrapped normal sample. Three outlier detection methods which are $M$, $A$ and $G_1$ statistics have been successfully applied on the data from wrapped normal distribution. The sampling behaviour and the performance of the test statistics are investigated via simulation. This study has showed that these three statistics perform well in identifying a single outlier that present in wrapped normal sample in certain cases by using simulation study. We illustrate the use of the statistics using the Kuantan wind direction data.

References

[1] Jammalamadaka S R and Sengupta A 2001 Topisc in Circular Statistics vol 5 Singapore: World Scientific Publishing Co. Pte. Ltd.
[2] Ravindran P and Ghosh S K 2011 Bayesian Analysis of Circular Data Using Wrapped Distributions. Journal of Statistical Theory and Practice 5(4) pp 547-561
[3] Kuppusamy M and Kaliyaperumal S K 2013 Comparison of methods for detecting outliers Scientific & Engineering Research 4(9)
[4] Hawkins D M 1980 Identification of Outliers Biometrical Journal 29(2) p 198
[5] Agostinelli C 2007 Robust estimation for circular data Computational Statistics & Data Analysis 51(12) pp 5867-5875
[6] Kaya A 2004 Outlier Effects on Database International Conference on Advances in Information Systems 3261
[7] Sun Z 2009 Comparing measures of fit for circular distributions Master thesis. University of Victoria
[8] Mardia K V and Jupp P E 1972 Directional Statistics London: John Wiley & Sons
[9] Collet D 1980 Outliers in Circular Data Journal of the Royal Statistical Society. Series C (Applied Statistics), 29(1) pp 50-57.
[10] Abuzaid A H, Mohamed I B and Hussin A G 2009 A new test of discordancy in circular data Communication in Statistics-Simulation and Computation 38(4) pp 682-691
[11] Mohamed I B, Rambli A, Khaliddin N and Ibrahim A I N 2015 A New Discordancy Test in Circular Data Using Spacings Theory Communications in Statistics Simulation and Computation

[12] Rambli A, Ibrahim S, Abdullah M I, Hussin A G and Mohamed I 2012 On Discordance Test for the Wrapped Normal Data Sains Malaysiana 14(6) pp 769-778

[13] Rambli A 2015 A half-circular distribution and outlier detection procedures in directional data Ph.D thesis. University of Malaya

[14] Sidik M I, Rambli A, Mahmud Z and Redzuan R S 2019 The Identification of Outliers in Wrapped Normal Data by using Ga Statistics International Journal of Innovative Technology and Exploring Engineering (IJITEE) 4(4) pp 181–188

[15] David HA 1970 Order Statistics New York and London: Wiley

[16] Barnett V and Lewis T 1984 Outliers in Statistical Data New York: John Wiley & Sons

Acknowledgments
Authors wishing to acknowledge assistance or encouragement from colleagues, special work by technical staff or financial support from organizations should do so in an unnumbered. We would like to extend our gratitude to the Centre of Statistical and Decision Sciences Studies, Faculty of Computer and Mathematical Sciences, Universiti Teknologi MARA (UiTM), Shah Alam. Lastly, we would also like to extend our gratitude to the Universiti Teknologi MARA Research Grants (600-IRMI/PERDANA 5/3 BESTARI (042/2018)) for providing the financial support in this study.