Abuzaid, Ali H.

Article

Detection of Outliers in Univariate Circular Data by Means of the Outlier Local Factor (LOF)

Statistics in Transition New Series

Provided in Cooperation with:
Polish Statistical Association

Suggested Citation: Abuzaid, Ali H. (2020) : Detection of Outliers in Univariate Circular Data by Means of the Outlier Local Factor (LOF), Statistics in Transition New Series, ISSN 2450-0291, Exeley, New York, Vol. 21, Iss. 3, pp. 39-51, https://doi.org/10.21307/stattrans-2020-043

STATISTICS IN TRANSITION new series, September 2020
Vol. 21, No. 3 pp. 39–51, DOI 10.21307/stattrans-2020-043
Received – 20.01.2020; accepted – 10.06.2020

Detection of Outliers in Univariate Circular Data by Means of the Outlier Local Factor (LOF)

Ali H. Abuzaid[1]

ABSTRACT

The problem of outlier detection in univariate circular data was the object of increased interest over the last decade. New numerical and graphical methods were developed for samples from different circular probability distributions. The main drawback of the existing methods is, however, that they are distribution-based and ignore the problem of multiple outliers.

The local outlier factor (LOF) is a density-based method for detecting outliers in multivariate data and it depends on the local density of every k nearest neighbours.

The aim of this paper is to extend the application of the LOF to the detection of possible outliers in circular samples, where the angles of circular data are represented in two Cartesian coordinates and treated as bivariate data. The performance of the LOF is compared against other existing numerical methods by means of a simulation based on the power of a test and the proportion of correct detection. The LOF performance is compatible with the best existing discordancy tests, while outperforming other tests. The level of the LOF performance is directly related to the contamination and concentration parameters, while having an inverse relationship with the sample size.

In order to illustrate the process, the LOF and other existing discordancy tests are applied to detect possible outliers in two common real circular datasets.

Key words: discordancy, distance, multiple outliers, neighbours, spacing theory.

1. Introduction

The analyses of directions in xy-plane is more convenient to be considered as circular data, which are distributed on a unit circle circumference, measured by degrees or radians and belonging to $\left[0^{\circ},360^{\circ}\right)$ or $\left[0,2\pi\right)$, respectively.

In the context of circular data, due to its closed bounded range property, then considering an outlier as an extreme value is no longer valid, where the extreme value

[1] Department of Mathematics, Al Azhar University – Gaza, Palestine. E-mail: a.abuzaid@alazhar.edu.ps. ORCID: https://orcid.org/ 0000-0002-6680-7371.

is defined as a point with the maximum circular deviation from the mean direction. Thus, the problem of outliers in circular data needs special discordancy tests. Collett (1980) proposed four tests of discordancy for circular samples. The past decade has seen a renewed interest in the detection and classification of outliers in the univariate circular data, either numerically (see Abuzaid et al. 2009, Mohamed 2016, Sidik et al. 2019) or graphically (Abuzaid et al. 2013). Recently, the problem of outliers in circular regression and functional relationship models has been well investigated (see, Satari et al. 2014, Alkasady et al. 2019).

Existing methods of outlier-detection in univariate circular data have some drawbacks: firstly, they are distribution-based methods, which rely on certain probabilistic distributional assumptions, where the cut-off points are needed for any combination of distribution parameters. Secondly, they were built for single outlier detection, and did not address the masking effect or multiple outliers. Lastly, they consider outlying as a binary property (i.e. either the angle is an outlier or not).

In geometrics, for a given angle θ with corresponding coordinates (x,y) on the unit circle, these coordinates are obtained as $x = \cos\theta$ and $y = \sin\theta$. Thus, treating the associated coordinates instead of the angle will allow us to use the available methods of outlier-detection in multivariate linear data. One of these methods is the local outlier factor (*LOF*), which is a density-based method. It computes the outlying factor of every point in a dataset based on its average distance to its k nearest neighbours. Furthermore, the outlier factor estimates the degree the suspected point is being outlying (Breunig et al. 2000). Recently, Abuziad (2020) has extended the concept of density-based local outliers to the medical multivariate circular data based on circular distances.

This article considers the LOF method, which is widely used in the multivariate analysis and available in most of statistical software programs as an alternative method of outlier-detection in univariate circular data, regardless of the probability distribution. The rest of this article is organized as follows: Section 2 reviews the main methods for outlier-detection in univariate circular data. Section 3 introduces the LOF in the circular data context. A comparative power of performance of available methods is presented in Section 4. For illustration, Section 5 analyses two real circular datasets.

2. Tests of discordancy in univariate circular data

Let $\theta_1,\ldots,\theta_n$ be a random sample from a circular variable, and the resultant length, $R = \sqrt{\left(\sum_{i=1}^{n} cos\,\theta_i\right)^2 + \left(\sum_{i=1}^{n} sin\,\theta_i\right)^2}$. The interest is to test the null hypothesis that θ_r, where

$1 \le r \le n$, is not an outlier. The following subsections review five discordancy tests to identify outliers.

2.1. M statistic

Mardia (1975) proposed a statistic based on the effect of removing the j th angle on the resultant length R, given by $M = \frac{R_r - R + 1}{n - R}$, where $R_r = max_j \left\{ R_{(-j)} \right\}$ and $R_{(-j)}$ is the resultant length after excluding the j th angle. The asymptotic distribution of M statistic is approximated by the standard normal distribution for large values of the concentration parameter (Collett, 1980).

2.2. C statistic

Collett (1980) proposed an alternative test of discordancy based on the mean resultant length, $\overline{R} = \frac{R}{n}$ and defined as $C = max_j \frac{\overline{R}_{(-j)} - \overline{R}}{\overline{R}}$, where $\overline{R}_{(-j)}$ is the mean resultant length after excluding the j th angle.

2.3. D statistic

The third statistic was derived by Collett (1980) based on the relative arc lengths between the ordered angles such as $\theta_{(1)} \le \theta_{(2)} \le ... \le \theta_{(n)}$. The arc length between consecutive angles is defined by $T_j = \theta_{(j+1)} - \theta_{(j)}, j = 1,...,n-1$ and $T_n = 2\pi - \theta_{(n)} + \theta_{(1)}$. The test statistic is given by $D_j = \frac{T_j}{T_{j-1}}, j = 1,...,n$. It corresponds to the greatest arc containing a single angle, θ_r, which is obtained by $D_r = \frac{T_r}{T_{r-1}}$. The $min\left\{ D_r, D_r^{-1} \right\}$ is considered because statistic D_r is a two-tailed statistic.

2.4. A statistic

Abuzaid et al. (2009) proposed a test statistic based on the summation of all circular distances from the angle θ_r to all other angles θ_j; $d_r = \sum_{j=1}^{n} 1 - cos\left(\theta_j - \theta_r \right)$ for j, $r = 1,...,n$. The test statistic is given by $max_r \left\{ \frac{d_r}{2(n-1)} \right\}, r = 1,...,n$. The approximated distribution of the A statistic was discussed in Abuzaid et al. (2012).

2.5. G statistic

Mohamed et al. (2016) extended the theory of arc length, which was used in D statistic to the spacing theory. The statistic is defined based on the a-step spacing where $a = 1,2,3,...,$ for the jth ordered angle as $G_{aj} = \theta_{(j+a)} - \theta_{(j)}$, for $j = 1,...,n-a$ and $G_{aj} = 2\pi - \theta_{(j)} + \theta_{(j+a)-n}$, for $j = (n+1)-a,(n+2)-a,...,n$. Then the test statistics is defined as $G_a = max_j \left\{ G_{aj} \right\}$.

To identify possible outliers in circular samples, the previous five test statistics have to exceed a pre-determined cut-off points which have been obtained via simulation under the assumptions that the circular data come from certain distribution with known sample size and parameters. The cut-off points and power of performance for the five statistics have been obtained for von Mises distribution and wrapped normal distribution (Sidik et al. 2019), while only the associated values of cut-off points for the first four statistics were obtained for the wrapped Cauchy distribution (Abuzaid et al. 2015) and Cardioid distribution (Das and Gogoi, 2015).

3. Local outlier factor (LOF) for univariate circular data

Breunig et al. (2000) proposed a density-based method for detecting outliers in multivariate data. It depends on the local density of every k nearest neighbours. It is the so-called a local outlier factor (*LOF*), and it does not consider the outlier as a binary property, where it assigns a factor for each point to indicate its outlying degree. The term "local" is derived from the fact that the value of the factor for a point θ depends on how that point is isolated with respect to the surrounding neighbourhood. A higher *LOF* value reflects more sparse neighbourhoods and represents an outlier point, while lower value of *LOF* reflects more dense neighbourhoods and represents a normal point.

LOF for a point θ is obtained by computing its average distance to its k nearest neighbours, then the distance is normalized by computing the average distance of each of those neighbours to their k nearest neighbours. The set of the following definitions explains the *LOF* algorithm.

1) Distance $d(\theta,\phi)$ between any two angles θ and ϕ:

Let θ and ϕ be two angles in a univariate circular dataset, φ, with coordinates of (x_θ, y_θ) and (x_ϕ, y_ϕ), respectively. Then, the distance between any two angles θ and ϕ is obtained by

$$d(\theta,\phi) = \sqrt{\left(x_\phi - x_\theta\right)^2 + \left(y_\phi - y_\theta\right)^2}\,.$$

2) k-distance of an angle θ:

For any positive integer k, the k-distance of an angle $\theta \in \varphi$ is denoted by $dist_k(\theta)$ and it is defined as the distance $d(\theta,\phi)$ between an angle θ and an angle $\phi \in \varphi$. It represents the k-th nearest neighbourhoods of an angle θ, where there is at least k angles such that $d(\theta,\nu) \le d(\theta,\phi)$ and at most $k-1$ angles, such that $d(\theta,\nu) < d(\theta,\phi)$, where ν is an angle and $\nu \in \varphi \backslash \{\theta\}$.

3) k-distance neighborhood of a point θ:

It contains every point whose distance from θ is not greater than the k-distance. It is defined as $N_k(\theta) = \{\phi \,|\, dist(\theta,\phi) \le dist_k(\theta)\}$ and it could be greater than k, where multiple points have the same distance.

4) Reachability distance from angle θ to angle ϕ:

For all close angles θ's to an angle ϕ, it is expected that there is a statistical fluctuation of $dist(\theta,\phi)$, which can be significantly reduced by defining the reachability distance as

$$reach\_dist_k(\theta,\phi) = max\{dist_k(\theta), d(\theta,\phi)\}\,.$$

The higher the value of k, the more similar the reachability distances for angles within the same neighbourhood.

5) Local reachability density of an angle θ:

It is the inverse of the average reachability distance based on the k nearest neighbours of an angle θ, and it is defined as

$$lrd_k(\theta) = \frac{N_k(\theta)}{\sum_{\phi \in N_k(\theta)} reach\_dist_k(\theta,\phi)}\,.$$

6) Local outlier factor of an angle θ:

It estimates the degree to which an angle θ is called an outlier, and it is defined as

$$LoF_k(\theta) = \frac{\sum_{\phi \in N_k(\theta)} \frac{lrd_k(\phi)}{lrd_k(\theta)}}{N_k(\theta)}.$$

It is the average of the ratio of the local reachability density of θ and those of θ's k-nearest neighbours.

The minimum number of neighbour angles to determine the density, which the so-called *MinPts*, and its effect on changing the values of the *LOF* was discussed by Breunig et al. (2000). They concluded that the *MinPts* can be between two and $n-1$, and suggested it to be at least 10 to remove unwanted statistical fluctuations. Furthermore, the angle is considered as an outlier if its *LOF* value is significantly greater than one.

In general A and G_a statistics have outperform other statistics (Sidik et al. 2019). Therefore, the following section will investigate the performance of the A statistic and *LOF* via simulation.

4. Power of performance

The performance of discordancy tests is evaluated by three measures, namely power function; $P1 = 1-\beta$ where β is the probability of type-II error, $P3$ which is the probability of identifying a contaminated value as an outlier when it is in fact an extreme value, and the probability of wrongly identifying a good observation as discordant, which is denoted by $P1-P3$ (Barnett and Lewis, 1984).

To obtain the three measures of performance, the following settings are considered in this simulation study, which were conducted based on 2000 random samples generated from two different circular distributions; namely the von Mises distribution with mean μ and concentration parameter κ; denoted as $vM(\mu,\kappa)$, and the wrapped Cauchy distribution with mean μ and concentration parameter ρ; denoted as $WC(\mu,\rho)$. Without loss of generality, the mean direction of generated samples from both distributions were fixed equal to zero. Five different sample sizes, namely $n = 20, 50, 70, 100$ and 150 were generated.

The considered values of concentration parameters are $\kappa = 0.5, 2, 5, 7$ and $\rho = 0.2, 0.4, 0.6, 0.8, 0.99$ for samples generated from von Mises and wrapped Cauchy distributions, respectively.

The samples are generated in such a way that $n-1$ of the observations come from the distribution, i.e. $vM\left(0,\kappa\right)$ or $WC\left(0,\rho\right)$, and the remaining one observation comes from $vM\left(\lambda\pi,\kappa\right)$ or $WC\left(\lambda\pi,\rho\right)$, respectively, where λ is the degree of contamination and $0\leq\lambda\leq1$. Then A statistic and *LOF* are calculated as given in Sections 2 and 3, respectively, where the value of k is fixed as the rounded up median for each sample size.

Figure 1 shows that *LOF* and A statistic are compatible in the case of samples from von Mises distribution, while *LOF* outperforms the A statistic in the case of samples from wrapped Cauchy distribution. The full results of the simulation study can be requested from the author. Simulation results show that two measures of performance, namely $P1$ and $P3$ are almost the same, thus the values of $P1-P3$ are always close to zero. The performance of outlier-detection methods is highly dependent on the circular distribution. In general, the methods of outlier-detection for samples from von Mises distribution perform significantly better than the case of wrapped Cauchy distribution. This may referred to the heavy tailed property of wrapped Cauchy distribution.

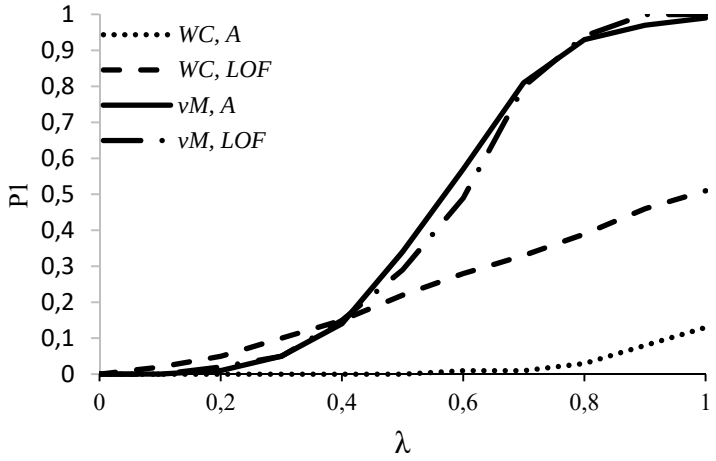

Figure 1. Performance of A statistic and *LOF*, for n =50, ρ =0.8 and κ=5

The performance of the *LOF* method has a direct relationship with the concentration parameter of circular sample as partially shown in Figure 2, while it has an inverse relationship with the sample size as shown in Figure 3. Moreover, in all considered cases, the performance has a direct relationship with the degree of contamination, λ.

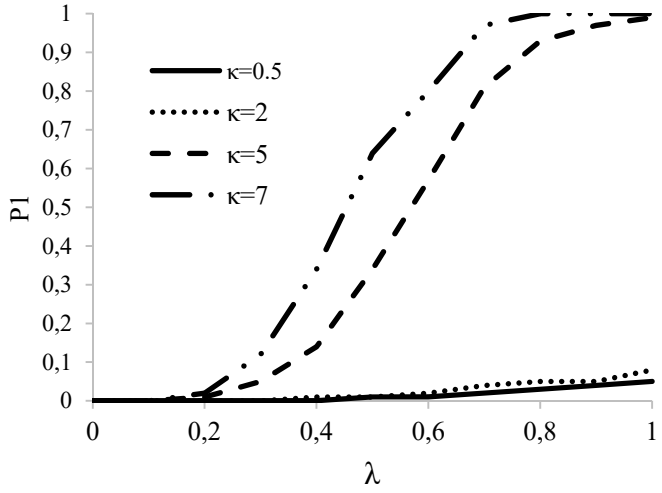

Figure 2. Performance of *LOF*, for samples of size n =50 from von Mises distribution

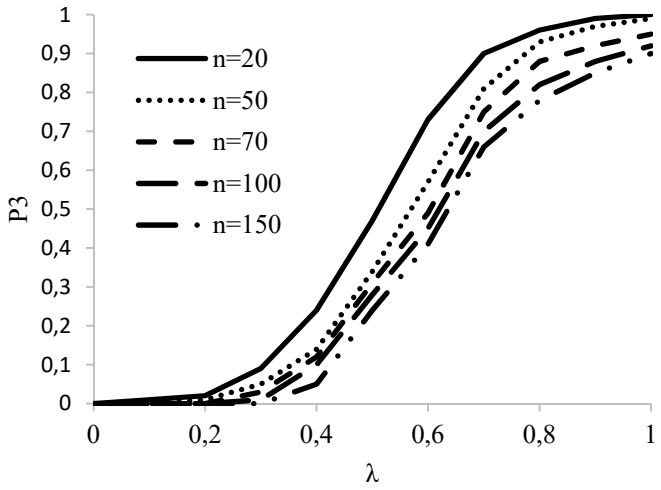

Figure 3. Performance of *LOF*, for samples with concentration parameter κ=5 from von Mises distribution

5. Practical examples

For illustration purposes, this section revisits two common circular datasets, which have been analysed in the context of outliers in circular data.

5.1. Frogs Data

Directions taken by 14 frogs after 30 hours of enclosure within a dark environmental chamber (Fergusion et al. 1967) are illustrated in Figure 4. The circular mean direction is 146.104° and the estimated concentration parameter is 2.18.

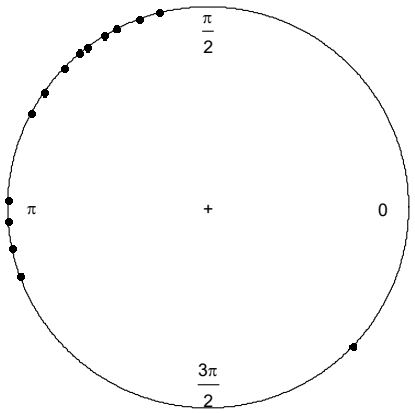

Figure 4. Circular plot of the frogs' directions, (n =14)

The results of applying seven discordancy tests on frogs' directions show that all tests except C statistic are consistent on identifying observation number 14 with value 316° (5.515 radians) as an outlier, which is apparent in Figure 4.

Table 1. Results of outlier-detection tests for frogs' directions, (n =14)

Statistic	*Statistic value*	*Observation*	*Cut-off point*	*Decision*
C	0.182	14	0.20	Not outlier
D	0.78	14	0.74	Outlier
M	0.52	14	0.50	Outlier
A	0.92	14	0.83	Outlier
G_1	2.03	14	1.69	Outlier
G_2	2.16	14	2.05	Outlier
LOF	1.88	14	1	Outlier

The values of *LOF* at k =10 are presented in Figure 5. It is shown that the *LOF* for all observations except the observation number 14 is close to one, which means that they are closed to each other and have similar density as their neighbours, while the *LOF* value of the observation number 14 is 1.88, which reveals that it has lower density than its neighbours.

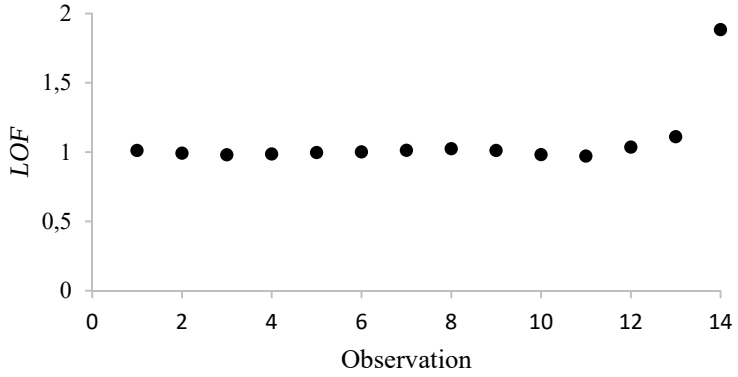

Figure 5. LOF for frogs' directions, k =10

5.2. Eye Data

Mohamed et al. (2016) considered the angle of posterior corneal curvature for 23 glaucoma patients as presented in Figure 6. The circular mean direction is 92° (1.61 radians) and the estimated concentration parameter is 6.84.

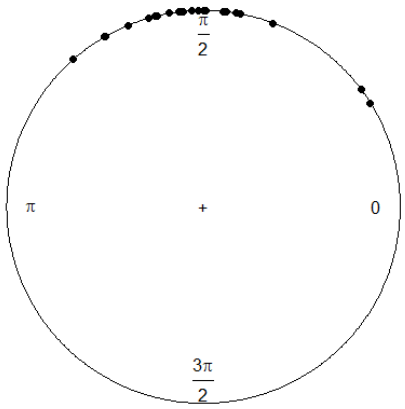

Figure 6. Circular plot of posterior corneal curvature, (n =23)

The results of applying discordancy tests on eye data show that only M statistic, G_2 and LOF at k =17 identified the observation number 17 as an outlier. Moreover, only G_2 and LOF identified the observation number 10 as an outlier. This reveals the weakness of most existing outlier-detection methods in the case of multiple outliers, which are apparently outliers from Figure 6.

Table 2. Results of outliers detection tests for eye dataset, (n =23)

Statistic	*Statistic value*	*Observation*	*Cut-off point*	*Decision*
C	0.02	17	0.03	Not outlier
D	0.04	17	0.18	Not outlier
M	0.31	17	0.12	Outlier
A	0.28	17	0.32	Not outlier
G_2	0.78	17	0.67	Outlier
G_2	0.68	10	0.67	Outlier
LOF	2.10	17	1	Outlier
LOF	1.97	10	1	Outlier

The values of LOF at k =17 are presented in Figure 7. It is shown that the LOF for all observations except the observation numbers 17 and 10 is close to one, which means that they are closed to each other and have similar density as their neighbours. On the other hand, the LOF values for the observation numbers 17 and 10 are 2.10 and 1.97, respectively, which reveals that they have lower density than their neighbours. Furthermore, the observation number 23 has a slightly high value of LOF and equals 1.36.

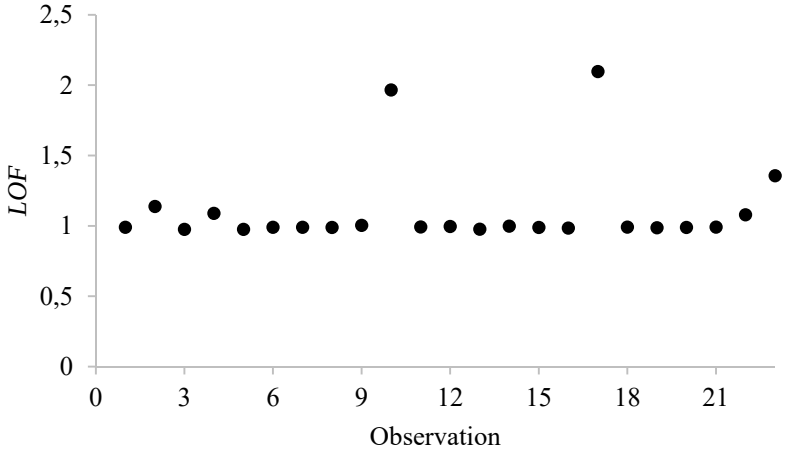

Figure 7. *LOF* for eye data, k =17

6. Conclusions

The presentation of angles in circular data as pairs of Cartesian coordinates gives a chance to use the *LOF* method for outlier detection. The *LOF* is a density-based method compared to the existing distribution-based methods. Furthermore, it does not

consider being an outlier as a binary property, while it gives the degree of being outlying.

The *LOF* performance is compatible with A test and then it outperforms the other tests of discordancy. The performance of the *LOF* has a direct relationship with the degree of contamination and concentration parameter, while it has an inverse relationship with the sample size.

The two considered practical examples illustrated the ability of *LOF* in dealing with multiple outliers compared to other existing outlier-detection methods.

The findings of this article pave the way to detect outliers in multivariate circular samples, either by representing their variates into pair coordinates, or by defining possible circular distances.

REFERENCES

ABUZAID, A. H. (2020). Identifying density-based local outliers in medical multivariate circular data. Statistics in Medicine. Vol. 39 (210), pp. 2793–2798. https://doi.org/10.1002/sim.8576.

ABUZAID, A. H., EL-HANJOURI, M. M., and KULLAB, M. M., (2015). On Discordance Tests for the Wrapped Cauchy Distribution, *Open Journal of Statistics*, Vol. 5 (4), pp. 245–253.

ABUZAID, A. H., MOHAMED, I. B., and HUSSIN, A. G., (2012). Boxplot for Circular Variables. *Computational Statistics*, Vol. 27 (3), pp. 381–392.

ABUZAID, A. H., RAMBLI, A., and HUSSIN, A.G., (2012). Statistics for a New Test of Discordance in Circular Data. *Communications in Statistics - Simulation and Computation*, Vol. 41 (10), pp. 1882–1890.

ABUZAID, A. M., MOHAMED, I. B., and HUSSIN, A. G., (2009). A new test of discordancy in circular data. *Communications in Statistics-Simulation and Computation,* Vol. 38(4), pp. 682–691.

ALKASADI, N. A., IBRAHIM, S. ABUZAID, A. H., and YUSOFF. M. I., (2019). Outliers Detection in Multiple Circular Regression Models Using *DFFITc* Statistic, *Sains Malaysiana*, Vol. 48 (7), pp. 1557–1563.

BARNETT, V. AND LEWIS, T., (1984). *Outliers in Statistical Data*. 2nd ed., John Wiley & Sons, Chichesters.

BREUNIG, M. M., KRIEGEL, H.-P., NG, R. T., and SANDER, JÖ., (2000). LOF: identifying density-based local outliers. *ACM sigmod record*, pp. 93–104.

COLLETT, D. (1980). Outliers in circular data. *Applied Statistics,* Vol. 29 (1), pp. 50–57.

DAS, M. K., GOGOI, B., (2015). Procedures of Outlier Detection in Cardioid Distribution, *Assam Statistical Review*, Vol. 29 (1), pp. 31–45.

FERGUSION, D. E., LANDRETH, H. F., MCKEOWN, J. P., (1967). Sun Compass Orientation of the Northern Cricket Frog. Acris Crepitans. *Animal Behavior,* Vol. 15, pp. 45–53.

MARDIA, K. V., (1975). Statistics of directional data. *Journal of the Royal Statistical Society, Series* B, Vol. 37, 349–393.

MOHAMED, I. B., RAMBLI, A., KHALIDDIN, N., and IBRAHIM, A. I. N., (2016). A New Discordancy Test in Circular Data Using Spacing's Theory, *Communications in Statistics - Simulation and Computation*, Vol. 45 (8), pp. 2904–2916.

SATARI, S. Z., HUSSIN, A. G, ZUBAIRI, Y. Z., and HASSAN, S. F., (2014). A New Functional Relationship Model For Circular Variables. *Pakistan Journal of Statistics*, Vol. 30 (3), pp. 397–410.

SIDIK, M. I., RAMBLI, A., MAHMUD, Z., REDZUAN, R. S., and SHAHRI, N., (2019). The Identification of Outliers in Wrapped Normal Data By Using Ga Statistics. *International Journal of Innovative Technology and Exploring Engineering*, Vol. 8 (4S), pp. 181–189.