Malaysian Journal of Science 38 (Special Issue 2): 46 - 54 (2019)

**THE INTERNATIONAL SEMINAR ON MATHEMATICS IN INDUSTRY (ISMI)
AND THE INTERNATIONAL CONFERENCE ON THEORETICAL AND APPLIED STATISTICS (ICTAS)
ISMI-ICTAS18 [4-6 SEPTEMBER 2018]**

An Outlier Detection Method for Circular Linear Functional Relationship Model Using *Covratio* Statistics

Nurkhairany Amyra Mokhtar[1a], Yong Zulina Zubairi[2b]*, Abdul Ghapor Hussin[1c] & Nor Hafizah Moslim[3d]

[1] Faculty of Defence Sciences and Technology, National Defence University of Malaysia, Kem Sungai Besi, 57000 Kuala Lumpur, MALAYSIA. E-mail: khairany.amyra@gmail.com[a] ; abdulghapor@gmail.com[c]
[2] Centre for Foundation Studies in Science,
University of Malaya, 50603 Kuala Lumpur, MALAYSIA. E-mail: yzulina@um.edu.my[b]
[3] Institute of Graduate Studies, University of Malaya, 50603 Kuala Lumpur, MALAYSIA. Email: moslimnorhafizah@gmail.com[d]
* Corresponding Author: yzulina@um.edu.my
Received: 21st April 2019 Revised : 6th August 2019 Published: 30th September 2019
DOI : https://doi.org/10.22452/mjs.sp2019no2.5

ABSTRACT The existence of outlier may affect data aberrantly. However, outlier detection problem has been frequently discussed for linear data but limited on circular data. Thus, this paper discusses an outlier detection method on circular data. We focus on circular data with equal error concentration parameters where the data is studied using linear functional relationship model. In this paper, the data and the error terms are distributed with von Mises distribution. We modify the *covratio* statistics in which the correction factor is applied to the estimation of concentration parameter. We develop the cut-off equation based on the 5% upper percentile of the *covratio* statistics and the power of performance of outlier detection is examined by a Monte Carlo simulation study. The simulation result shows that the power of performance increases when the concentration and the level of contamination increase. The applicability of the proposed method is illustrated by using the wind direction data collected from the Holderness Coastline at the Humberside Coast in North Sea, United Kingdom.

Keywords: circular data, linear functional relationship model, outlier detection, *covratio* statistics

1. INTRODUCTION

Circular observation is specified by the angle from the initial direction to the point on the circle which corresponds to the observation, after an initial direction and an orientation of the circle have been chosen. It is measured in degrees or radians (Mardia & Jupp, 2000). Some examples of circular data include the orientation of fracture planes, the wind direction and the direction of the ocean current (Fisher, 1993).

The most useful distribution on the circle is said to be the Von Mises distribution (Mardia & Jupp, 2000). The probability density function of the distribution is

$$g(\theta;\mu,\kappa)=\frac{1}{2\pi I_0(\kappa)}e^{\kappa\cos(\theta-\mu)} \quad (1)$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero, which can be defined by:

$$I_0(\kappa)=\frac{1}{2\pi}\int_0^{2\pi}e^{\kappa\cos\theta}d\theta \quad (2)$$

for $0\le\theta<2\pi$ and $\kappa>0$ where μ is the mean direction and κ is the concentration parameter.

Functional relationship model for

circular variables may be used to compare circular data where both of the variables are subjected to errors (Hassan et al., 2010). In this model, both X and Y variables are subject to random errors δ_i and ε_i, respectively. The errors are distributed with Von Mises distribution of $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,\kappa)$. Caires and Wyatt (2003) developed a model with the desired symmetry of the functional relationship model. The model is named the linear functional relationship model and is written in the form

$$Y = \alpha + X \left(\mathrm{mod}\, 2\pi\right) \tag{3}$$

where α is the rotation parameter.

Outlier detection has important applications such as fraud detection and robust analysis, among others (Mokhtar et al., 2018). *Covratio* statistics have been developed for linear data (Ghapor et al., 2014). However, for circular data, the method is somewhat limited, especially for linear functional relationship model. Therefore, this paper discusses the *covratio* statistics used for outlier detection in a linear functional relationship model for circular variables.

Section 2 describes the method used to obtain the cut-off equation to detect the outlier. Section 3 shows the results of the power of performance of the cut-off equation obtained in detecting the outlier. Section 4 illustrates the applicability of the proposed method and the conclusion is deduced in Section 5.

2. METHODS

2.1 Maximum Likelihood Estimation of Parameters of Von Mises Distribution

The log-likelihood function of the Von Mises distribution, based on observations x and y is given by

$$\log L(\alpha,\kappa,\nu,X;x,y) = -2n\log 2\pi - n\log I_0(\kappa) - n\log I_0(\nu) + \kappa\sum_{i=1}^{n}\cos\left(x_i - X_i\right) + \nu\sum_{1}^{n}\cos\left(y_i - \alpha - X_i\right) \tag{4}$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero and κ is the concentration parameter.

In this case, the ratio of concentration parameter $\lambda = \frac{\nu}{\kappa}$ is fixed as 1 thus $\kappa = \nu$. Therefore, the log-likelihood function of the Von Mises distribution then becomes

$$\log L(\alpha,\kappa,X;x,y) = -2n\log 2\pi - 2n\log I_0(\kappa) + \kappa\sum_{i=1}^{n}\cos\left(x_i - X_i\right) + \kappa\sum_{i=1}^{n}\cos\left(y_i - \alpha - X_i\right) \tag{5}$$

The estimate of the rotation parameter for this LFRM is then

$$\hat{\alpha} = \begin{cases} \tan^{-1}\left\{\dfrac{S}{C}\right\} & \text{when } S>0, C>0 \\ \tan^{-1}\left\{\dfrac{S}{C}\right\}+\pi & \text{when } C<0 \\ \tan^{-1}\left\{\dfrac{S}{C}\right\}+2\pi & \text{when } S<0, C>0 \end{cases} \tag{6}$$

where $S=\sum_{i=1}^{n}\sin\left(y_i-\hat{X}_i\right)$ and $C=\sum_{i=1}^{n}\cos\left(y_i-\hat{X}_i\right)$. Meanwhile,

$$\hat{\kappa}=A_1^{-1}\left(\frac{1}{n}\left\{\sum_{i=1}^{n}\cos\left(x_i-\hat{X}_i\right)+\sum_{i=1}^{n}\cos\left(y_i-\alpha-\hat{X}_i\right)\right\}\right) \tag{7}$$

$A_1(x)$ is a function that behaves rather like $\left(\frac{2}{\pi}\right)\tan^{-1}x$ and so $A_1^{-1}(x)$ is like $\tan\left(\frac{\pi x}{2}\right)$ (Dobson, 1978). In this model, we use the approximation for the estimation of the concentration parameter κ that has been given by Fisher (1993) for the case of equal error concentration as a piecewise function of :

$$A_1^{-1}(w)=\begin{cases} 2w+w^3+\dfrac{5}{6}w^3 & \text{when } w<0.53 \\ -0.4+1.39w+\dfrac{0.43}{(1-w)} & \text{when } 0.53\le w<0.85 \\ \dfrac{1}{w^3-4w^2+3w} & \text{when } w\ge 0.85 \end{cases} \tag{8}$$

where

$$w=\frac{1}{n}\left\{\sum_{i=1}^{n}\cos\left(x_i-\hat{X}_i\right)+\sum_{i=1}^{n}\cos\left(y_i-\hat{\alpha}-\hat{X}_i\right)\right\}.$$

Caires and Wyatt (2003) noted that, in the circular case, the estimation of a concentration parameter (whose inverse is equivalent to the variance for linear data) needs to be corrected by dividing it by 2 as it suggests a consistent estimator of κ (Caires & Wyatt, 2003). It has been proposed that $\tilde{\kappa}=\frac{\hat{\kappa}}{2}$ gives a better approximation to the value of κ.

X_i is solved iteratively by some initial guess. Suppose $\hat{X}_{i0}$ is an initial estimate of $\hat{X}_i$. Then $x_i-\hat{X}_i=x_i-\hat{X}_{i0}+\hat{X}_{i0}-\hat{X}_i=\left(x_i-\hat{X}_{i0}\right)+\Delta_i$ where $\Delta_i=\hat{X}_{i0}-\hat{X}_i$. Thus, $y_i-\hat{\alpha}-\hat{X}_i=(y_i-\hat{\alpha}-\hat{X}_{i0})+\Delta_i$. Hence, the partial derivative equation above becomes: $\sin\left(x_i-\hat{X}_{i0}+\Delta_i\right)+\sin(y_i-\hat{\alpha}-\hat{X}_{i0}+\Delta_i)=0$

when Δ_i is small, then $\cos\Delta_i \approx 1$ and $\sin\Delta_i \approx \Delta_i$. Therefore, the variable X may be estimated by:

$$\hat{X}_{i1} \approx \hat{X}_{10} + \frac{\sin\left(x_i - \hat{X}_{i0}\right) + \sin\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)}{\cos\left(x_i - \hat{X}_{i0}\right) + \cos\left(y_i - \hat{\alpha} - \hat{X}_{i0}\right)} \tag{9}$$

2.2 Covratio Statistics to Detect Outlier in the LFRM for Circular Variables Assuming Equal Error Concentration Parameters

The outlier is an observation that lies outside the pattern (Belsley et al., 1980). Various outlier detections have been employed. The *covratio* statistics have long been used to identify outlier in linear regression models via a row deletion approach. Some studies on detecting outliers have been discussed by Abuzaid et al. (2011), Ibrahim et al. (2013), Ghapor et al. (2014) and Rambli et al. (2016). The *covratio* statistic is used to measure the effect of removing the observation based on the determinantal ratio given by:

$$COVRATIO_{(-i)} = \frac{|COV|}{|COV_{(-i)}|} \tag{10}$$

where $|COV|$ is the determinant of covariance matrix for the full set and $|COV_{(-i)}|$ is the determinant of covariance matrix for the reduced data set by excluding the *i-th* row.

In 2015, Mokhtar et al. (2015) studied the LFRM with the assumption of equal error concentration parameter and the covariance matrix of the parameter in the model is as given.

$$\operatorname{cov}\begin{bmatrix}\hat{\alpha}\\ \tilde{\kappa}\end{bmatrix} = \begin{bmatrix}\frac{1}{2nA_1'(\tilde{\kappa})} & 0\\ 0 & \frac{2}{n\tilde{\kappa}A_1'(\tilde{\kappa})}\end{bmatrix} \tag{11}$$

where $\operatorname{var}(\hat{\alpha}) = \frac{1}{2nA_1'(\tilde{\kappa})}$ and $\operatorname{var}(\tilde{\kappa}) = \frac{2}{n\tilde{\kappa}A_1'(\tilde{\kappa})}$, and $A_1'(\tilde{\kappa}) = 1 - \frac{A_1(\tilde{\kappa})}{\tilde{\kappa}} - [A_1(\tilde{\kappa})]^2$ is the first derivative $\frac{dA_1(\tilde{\kappa})}{d\tilde{\kappa}}$ for the function $A_1(\tilde{\kappa})$ which is the ratio of first and zeroth order Bessel functions.

Here, we propose that the *covratio* statistics method in which the correction factor is applied to estimate the concentration parameter. Therefore, the determinant of the covariance matrix for this model becomes

$$|COV| = \frac{1}{n^2\tilde{\kappa}[A_1'(\tilde{\kappa})]^2} \tag{12}$$

The observation with $\left|COVRATIO_{(-i)}-1\right|$ that exceeds the cut-off points will be identified as an outlier. The steps in determining the cut-off point are discussed in the next section.

2.3 Simulation Study to Determine the Cut-Off Equation using Covratio Statistics

In detecting the outlier, a cut-off point is needed as an indicator to examine the power of performance of the proposed *covratio* statistics. Therefore, a Monte Carlo simulation study is performed with different values of sample size and error concentration parameter. In this part, the number of simulation $s=500$.

Without the loss of generality, the variable X is generated from the Von Mises distribution of $VM(2,3)$ and the value of $\alpha=\frac{\pi}{4}=0.7854$. The values of the concentration parameters of the error term used in this study are $\kappa=3,5,10$ and 15. For each value of κ, the sample size $n=20,\ 30,\ 50,\ 70,\ 100,130$ and 150 are considered for the simulation with the assumption of $\kappa=\nu$.

The process was repeated for 500 simulations and the 5% upper percentiles of the maximum $\left|COVRATIO_{(-i)}-1\right|$ were obtained. The values of the upper 5% percentiles were then used to construct a cut-off equation in identifying the outlier for the LFRM for equal error concentration parameters. Table 1 shows the values of 5% upper percentile to consider the 95% confidence level.

Table 1: The values of the upper 5% percentile of $\left|COVRATIO_{(-i)}-1\right|$

n	$\kappa=3$	$\kappa=5$	$\kappa=10$	$\kappa=15$
20	0.482631	0.456601	0.427134	0.440119
30	0.345093	0.341747	0.327810	0.325791
50	0.247297	0.245086	0.213416	0.217016
70	0.194456	0.204701	0.162582	0.163800
100	0.134861	0.201583	0.126035	0.119783
130	0.113857	0.161380	0.104580	0.097920
150	0.101447	0.147530	0.091868	0.083569

The arithmetic mean of the values for the respected n are calculated and by finding the best fit of using least square method and thus a power function is obtained. Thus, the fit of the cut-off equation is obtained with $y=3.7586n^{-0.71}$. Figure 1 shows the plot of the best fit and the power fit has a good fit of R^2 almost equal to 1.

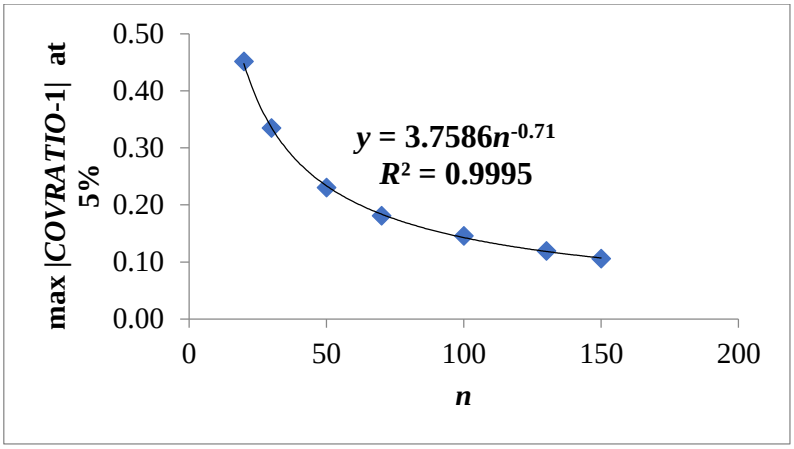

Figure 1: The power series graph to determine the cut-off equation at 5% significance level for equal error concentration

2.4 *Simulation Study to Assess the Power of Performance of Covratio Statistics in Outlier Detection*

To assess the feasibility of the proposed method, the power of performance is examined through a Monte Carlo simulation study with a number of simulations, $s = 500$. The steps below were carried out to identify the power of performance of the proposed *covratio* statistics in detecting the outlier.

Step 1: The values of X variable were generated from the distribution $VM(2,3)$ with the size of $n = 30, 70, 100$ and 130 and $\kappa = 10, 15$ and 20. An observation X_d^* is then contaminated with some levels of contamination ω where the level of the contamination was $0 \le \omega \le 1$ using the formula $X_d^* = X_i + \omega\pi \pmod{2\pi}$.

Step 2: Values of Y were found according to the generated X values. The variables X and Y were considered with the generated random error terms of $\delta_i \sim VM(0,\kappa)$ and $\varepsilon_i \sim VM(0,\nu)$, respectively where $\kappa = \nu$.

Step 3: The variables were fitted to the LFRM and the concentration parameter was estimated with the correction factor mentioned in Section 2.1.

Step 4: Calculate the value of $COVRATIO_{(-i)}$ and find $\left|COVRATIO_{(-i)} - 1\right|$ for all i. If the value of $\left|COVRATIO_{(-i)} - 1\right|$ exceeded $y = 3.7586n^{-0.71}$, then the ith observation is marked as a contaminated observation.

Step 5: The percentage of correct detection of the outlier was calculated as the power of performance.

3. RESULTS

Table 2 shows the simulation results in assessing the power of performance of $\left|COVRATIO_{(-i)} - 1\right|$ in LFRM for circular variables assuming equal error concentration parameters. The percentage of correct detection of the outlier is calculated when the maximum value of $\left|COVRATIO_{(-i)} - 1\right|$ exceeds $y = 3.7586n^{-0.71}$.

Table 2: The power of performance of $\left|COVRATIO_{(-i)} - 1\right|$

n	ω	κ =10	κ =15	κ =20
30	**0.2**	4.60	6.40	9.20
	0.4	21.60	41.80	60.60
	0.6	66.20	85.00	99.00
	0.8	94.20	99.80	100.00
	1.0	99.20	100.00	100.00
70	**0.2**	1.80	3.40	6.00
	0.4	14.20	31.20	54.60
	0.6	64.00	86.40	97.40
	0.8	93.60	99.60	100.00
	1	99.40	100.00	100.00
100	**0.2**	2.80	1.60	4.20
	0.4	10.80	33.00	45.80
	0.6	64.00	86.40	97.40
	0.8	92.20	99.20	100.00
	1	98.80	100.00	100.00
130	**0.2**	1.00	1.20	2.60
	0.4	10.20	25.00	46.40
	0.6	49.20	84.20	96.80
	0.8	88.00	99.40	100.00
	1	98.60	100.00	100.00

As an example, Figure 2 shows the plot of the power of performance for $n = 130$.

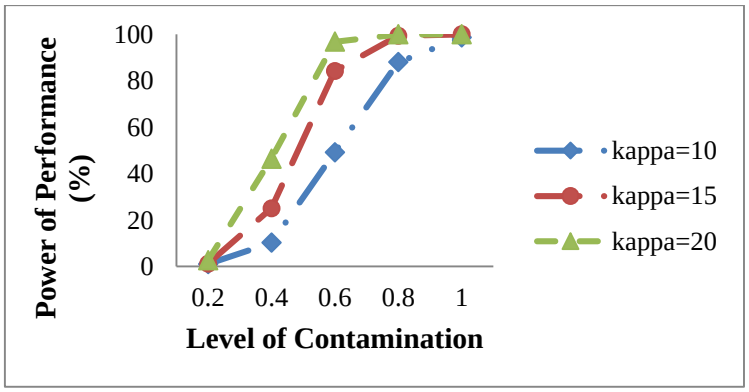

Figure 2: Power of performance for *covratio* statistics in detecting outlier for $n = 130$

The simulation study shows that the power of performance increases as the concentration parameter and the level of contamination increase. The highest concentration parameter with the highest level of contamination gives the highest power of performance where the value is 100%. Therefore, the *covratio* statistic method used in this study, in which the covariance matrix is derived with some correction factor for the maximum likelihood estimation, is sufficient to detect outlier in circular data with error terms.

4. APPLICATION ON REAL WIND DIRECTION DATA

The proposed method is applied to a real wind direction data set with sample size of $n = 129$ obtained from Holderness Coastline, Humberside Coast, United Kingdom (Hussin et al., 2004). The data is collected over the period of 22.7 days. Variable *x* is the data of the wind direction measured by high frequency radar system and developed by UK Rutherford and Appleton Laboratories, using the pulse radar operating at frequency of 24.2-27 MHz. Meanwhile for the variable *y*, the data was measured using an anchored wave buoy. Previous researchers of circular statistics such as Mokhtar et al. (2018) and Hussin et al. (2013) have used this data to illustrate the presence of outliers. It is worthwhile to note that the values of error concentration parameters of the variables *x* and *y* are assumed as equal. They have established that observations 38 and 111 as outliers of the data set.

Figure 3 shows the scatterplot of the values of $\left|COVRATIO_{(-i)} - 1\right|$ for all 129 observations in the data.

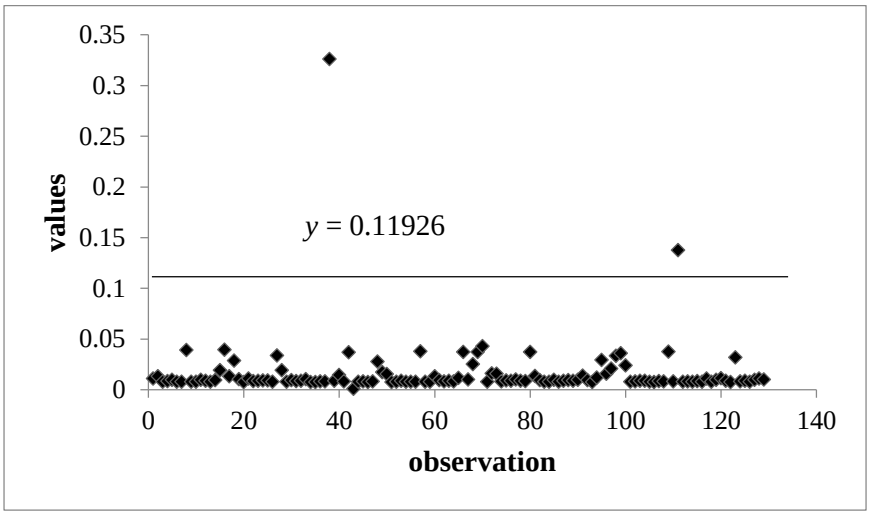

Figure 3: Values of $\left|COVRATIO_{(-i)} - 1\right|$ for all 129 observations of the real wind direction data.

Based on the scatterplot, it can be seen that the values of $\left|COVRATIO_{(-i)} - 1\right|$ for observations 38 and 111 exceeded the cut-off equation $y = 3.7586n^{-0.71} = 0.11926$. Therefore, observations 38 and 111 were detected as the outliers of the data set. The value of $\left|COVRATIO_{(-i)} - 1\right|$ for these two observations exceed the cut-off equation $y = 3.7586n^{-0.71}$.

5. CONCLUSION

This paper proposed an outlier detection method for the linear functional relationship model of circular variables with equal error concentration parameters. The covariance matrix is derived with some correction factor applied to the maximum likelihood estimation to obtain the *covratio* statistics of the model. This study considers a 95% confidence level, and the cut-off equation developed is to be at 5% significant level, read as $y = 3.7586n^{-0.71}$. The pattern in the power of performance in the simulation study shows that this method is adequate to detect the outlier that exists in a circular data.

6. ACKNOWLEDGEMENT

We would like to thank National Defence

University of Malaysia, University of Malaya (Grant GPF006H-2018), the Ministry of Education Malaysia and GE STEM grant (vot no. 07397) for supporting this work.

7. REFERENCES

Abuzaid, A., Mohamed I, Hussin, A. G. and Rambli, A. (2011). Covratio statistics for simple circular regression model. *Chiang Mai Journal of Science*, 2011. 38 (3): 321-330.

Belsley, D. A., Kuh, E and Welsch, R. E. (1980). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. New York: Wiley.

Caires, S. and Wyatt, L. R. (2003). A linear functional relationship model for circular data with an application to the assessment of ocean wave measurement. *Journal of Agricultural, Biological, and Environmental Statistics, Biological and Environmental Statistics*, 8 (2): 153-169.

Dobson, A. (1978). Simple approximations for the Von Mises concentration statistic. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 27(3): 345-347.

Duan, L., Xu, L., Liu, Y. and Lee, J. (2009). Cluster-based outlier detection. *Annals of Operations Research*, 168(1): 151-168.

Fisher, N. I. (1993). *Statistical Analysis of Circular Data*. United Kingdom: Cambridge University Press.

Ghapor, A. A, Zubairi, Y. Z, Mamun, A. S. M. A. and Imon, A. H. M. R. (2014). On detecting outlier in simple linear functional relationship model using *covratio* statistic. *Pakistan Journal of Statistics*, 30(1): 129-142.

Hassan, S. F, Hussin, A. G. and Zubairi, Y. Z. (2010). Estimation of functional relationship model for circular variables and its application in measurement problem. *Chiang Mai Journal of Science*, 37(2): 195-205.

Hussin, A. G., Fieller, N. R. J. and Stillman, E. C. (2004). Linear regression for circular variables with application to directional data. *Journal of Applied Science & Technology*, 8(1 & 2): 1-6.

Hussin, A.G., Abuzaid, A.H., Ibrahim, A.I.N. & Rambli, A. (2013). Detection of outliers in the complex linear regression model. *Sains Malaysiana*, 42(6): 869–874

Ibrahim S, Rambli A, Hussin A G and Mohamed I. (2013). Outlier detection in a circular regression model using covratio statistic. *Communication in Statistics-Simulation and Computation*, 42(10): 2272-2280.

Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics*. New Jersey: John Wiley & Sons.

Mokhtar, N. A., Zubairi, Y. Z. and Hussin, A. G. (2015). A simple linear functional relationship model for circular variables and its application. *Proceedings of the 9th International Conference on Renewable Energy Sources (RES '15)*, Kuala Lumpur, Malaysia, pp. 57-63.

Mokhtar, N. A., Zubairi, Y. Z., & Hussin, A. G. (2018). A clustering approach to detect multiple outliers in linear functional relationship model for circular data. *Journal of Applied Statistics*, 45(6): 1041-1051.

Rambli, A., Abuzaid, A. H. M. , Mohamed, I. B. and Hussin, A. G. (2016). Procedure for detecting outliers in a circular regression model. *PLOS ONE*, 11(4): e0153074.