DIGITALCOMMONS
— @WAYNESTATE —

Journal of Modern Applied Statistical Methods

Volume 16 | Issue 2 Article 22

December 2017

Detection of Outliers in Univariate Circular Data using Robust Circular Distance

Ehab A. Mahmood
University Putra Malaysia, Serdang, Malaysia, eee.mahmood@gmail.com

Sohel Rana
East West University, Dhaka, Bangladesh, srana_stat@yahoo.com

Habshah Midi
University Putra Malaysia, Serdang, Malaysia, habshahmidi@gmail.com

Abdul Ghapor Hussin
National Defence University of Malaysia, Kuala Lumpur, Malaysia, abdulghapor@gmail.com

Recommended Citation

Mahmood, E. A., Rana, S., Midi, H., & Hussin, A. G. (2017). Detection of Outliers in Univariate Circular Data using Robust Circular Distance. Journal of Modern Applied Statistical Methods, 16(2), 418-438. doi: 10.22237/jmasm/1509495720

Journal of Modern Applied Statistical Methods
November 2017, Vol. 16, No. 2, 418-438.
doi: 10.22237/jmasm/1509495720

ISSN 1538 − 9472

Detection of Outliers in Univariate Circular Data using Robust Circular Distance

Ehab A. Mahmood
University Putra Malaysia
Serdang, Malaysia

Sohel Rana
East West University
Dhaka, Bangladesh

Habshah Midi
University Putra Malaysia
Serdang, Malaysia

Abdul Ghapor Hussin
National Defence University of Malaysia
Kuala Lumpur, Malaysia

A robust statistic to detect single and multi-outliers in univariate circular data is proposed. The performance of the proposed statistic was tested by applying it to a simulation study and to three real data sets, and was demonstrated to be robust.

Keywords: Circular data, outliers, masking, swamping

Introduction

Statistical data can be classified according to their distributional topologies into two sets. First, linear data can be represented on a straight line. Second, circular data can be represented on the circumference of a unit circle. Circular data can be measured either in degrees, when they are distributed in the interval [0° – 360°), or in radians, in the interval [0 – 2π). Circular data can arise in contrasting scientific fields such as earth sciences, meteorology, biology, physics, psychology, image analysis, and medicine. The classical statistics that we apply to linear data cannot be used for circular data because of the geometrical properties of the circular data. For example, if we have two circular data points at 100° and 300°, then the arithmetic mean according to the linear measure is equal to 200°. However, the mean direction is equal to 80° according to the geometrical theory of the circle.

A special statistical measure is needed to deal with circular data. Between 5% and 10% of any set of statistical data are surprising points that are often called outliers (Hampel, Ronchetti, Rousseeuw, & Stahel, 1986). They may unduly affect the statistical analysis and the final outcomes.

Ehab A. Mahmood is a PhD student in the Department of Mathematics, Faculty of Science. Email them at: eee.mahmood@gmail.com.

There are many methods to detect outliers in linear univariate data. However, little is known regarding detecting outliers in circular data. It can be defined as an observation that is discordant in comparison with the rest of the sample. In linear data, an outlier is an observation that is extreme. However, the outliers in circular data may not have extreme values. For example, consider the following circular data (Collett, 1980):

$$10, 15, 20, 25, 30, 350$$

As linear data, it is obvious that the last observation, 350, is an outlier. However, as circular data, it is clear that 350 is consistent with the other observations. Therefore, the methods used to detect outliers in circular data are different from those used for linear data.

Mardia (1975) suggested a statistic to identify a single outlier in univariate circular data. He considered the observation that is the most influential on the resultant length to be an outlier. Collett (1980) proposed four test statistics, namely *L*, *C*, *D*, and *M*, to identify a single outlier in univariate circular data. It was found for small samples sizes that it is better to use the *C* and *D* statistics. However, no statistic was recommended to detect multiple outliers, and typical methods are only successful in detecting a single outlier. Furthermore, there was no discussion on how to identify an outlier when the sample size is large.

Fisher (1993) summarized three causes of outliers in statistical data: mis-recording, unwitting sampling from another population, and vagaries of sampling resulting in the occasional isolated value. In this identification he used the *M* statistic, which had already been suggested by Collett (1980), and did not propose a new statistic.

Mardia and Jupp (2000) suggested that circular data could be tested by considering three factors: The first was the mean resultant length. They promoted the use of either the *Mardia* (Mardia, 1975) statistic or the *C* statistic (Collett, 1980). The second was the likelihood ratio test for slippage in the model. For circular data, they considered either the likelihood ratio test for location slippage in a von Mises distribution (Collett, 1980) or the likelihood ratio test for concentration slippage in a Fisher distribution (Fisher, Lewis, & Willcox, 1981). Their final factor was the exponential distribution. Some tests for this factor had been suggested by Fisher et al. However, Mardi and Jupp did not suggest a new test statistic.

Jammalamadaka and SenGupta (2001) promoted the use of the *P-P* plot as a simple graphical way of detecting outliers in circular data. Furthermore, they proposed two statistics: The first of these was the locally most powerful invariant

(LMPI) statistic. They used LMPI for the circular data that were applie din the wrapped stable and uniform mixture model WSM. Second, they proposed using a likelihood ratio testing (LRT) approach to identify outliers in circular data. They tested the hypothesis that $\vartheta_1, \vartheta_2, \ldots, \vartheta_{i-1}, \vartheta_{i+1}, \ldots, \vartheta_n$ follow a von Mises distribution [vM(μ_0, k)] and that ϑ_i is distributed as [vM(μ_1, k)], where i is unknown. They applied this test to two cases: first, when μ_0, μ_1, and k are all known and second, when only k is known. They calculated the power of the procedure and the probability of detecting outliers perfectly by comparing both the LMPI and the LRT approaches with the L statistic (Collett, 1980) and with the statistic of Mardia (1975). They noted that their statistics are better than the other statistics. However, they did not propose a way to test circular data if k is unknown.

Otieno and Anderson-Cook (2005) tested three of the preferred directions, mean direction, median direction (Fisher, 1993), and the Hodges-Lehmann (HL) estimate (Otieno & Anderson-Cook, 2003). They concluded circular HL is a good compromise between circular mean and circular median, like its counterpart for linear data. The HL estimator is less robust to outliers compared to the median, but it is an efficient alternative because it has a smaller circular variance.

Abuzaid, Mohamed, and Hussin (2009) proposed the A statistic to detect an outlier in univariate circular data. This depends on the sum of the circular distances from any point to all other points on the circumference of the unit circle. They depended on calculating both the probability that the contaminant observation was an extreme observation and could be identified as an outlier and the probability of a type II error as a measure for comparing their suggestion with the C, D, and M statistics. However, the probability results for the A statistic are close to the results for the C statistic and they did not test their suggestion for large sample size or apply it to identify multiple outliers.

Abuzaid (2010) used the geometrical properties of the chord of a circle for detecting an outlier in univariate circular data. However, this suggestion detects only a single outlier. Abuzaid, Hussin, Rambli, and Mohamed (2012) then suggested a test statistic to detect outliers in univariate and bivariate circular data. The test statistic was based on the approximate distribution of the circular distances between the sample points. Nonetheless, they did not evaluate their statistic using any statistical measures. Moreover, they suggested a way to identify only a single outlier.

Mohamed, Rambli, Khaliddin, and Ibrahim (2015) proposed a procedure to identify single outliers and patches of outliers in univariate circular data. It is based on spacing theory in circular data. They compared their procedure with the C, D, and A statistics. However, their procedure is difficult, especially if the circular data have multiple outliers. Furthermore, the rates of swamping, the identification of inliers as outliers (Barnett & Lewis, 1978), are relatively high. For more information, see Beckman and Cook (1983) and Barnett and Lewis (1978), who reviewed the literature on the detection of outliers in various areas of statistical data.

The aim of this study is to investigate the robustness of our proposed statistic to detect outliers in univariate circular data when the data follow a von Mises distribution. The circular distance between any circular data point and the circular median is considered as a statistic test to identify outliers.

The circular median is defined as any angle ϑ such that half of the data points lie in the arc ϑ, $\vartheta + \pi$ and the majority of the data points are nearer to ϑ than to $\vartheta + \pi$ (Mardia & Jupp, 2000). This will be compared with existing methods to detect a single outlier. Furthermore, the aim is to identify outliers when there is a high level of contamination and with large sample sizes, using various statistical measures to evaluate the procedure.

von Mises Distribution

Let ϑ_1, ϑ_2,…, ϑ_n be circular observations following a von Mises distribution with mean direction μ and concentration parameter k, denoted by [vM(μ, k)]. The probability density function of the von Mises distribution is given by Hamelryck, Mardia, and Ferkinghoff-Borg (2012) as follows:

$$\mathrm{g}\left(\vartheta,\mu,k\right)=\frac{1}{2\pi\,\mathrm{I}_0}\mathrm{e}^{k\cos\left(\vartheta-\mu\right)} \tag{1}$$

where $0 \leq \mu < 2\pi$, $k \geq 0$, and I_0 denotes the modified Bessel function of the first kind and order 0, which can be defined as follows:

$$\mathrm{I}_0\left(k\right)=\frac{1}{2\pi}\int_0^{2\pi}\mathrm{e}^{k\cos\left(\vartheta\right)}d\vartheta$$

If $k = 0$, then the probability density function of the von Mises distribution will be the same as the probability density function of the uniform distribution of circular data, where

$$\mathrm{g}\left(\vartheta,\mu,0\right)=\frac{1}{2\pi}$$

The mean direction of the circular observations is estimated according to the following formula (Fisher, 1993):

$$\hat{\mu}=\begin{cases}\tan^{-1}\left(s/c\right) & \text{if } s>0, c>0\\ \tan^{-1}\left(s/c\right)+\pi & \text{if } c<0\\ \tan^{-1}\left(s/c\right)+2\pi & \text{if } s<0, c>0\end{cases} \tag{2}$$

where

$$s=\sum_{i=1}^{n}\sin\left(\vartheta_i\right),\quad c=\sum_{i=1}^{n}\cos\left(\vartheta_i\right)$$

The mean resultant length $\bar{R}$ is a measure of the concentration of the circular observations at a specific point of the circumference of the circle. It is calculated using this formula:

$$\bar{R}=\sqrt{\bar{c}^2+\bar{s}^2} \tag{3}$$

where $0 \le \bar{R} \le 1$, $\bar{c} = c / n$, $\bar{s} = s / n$.

$\bar{R} = 0$ is satisfied if and only if the circular data are widely dispersed on the circumference ($\bar{c} = 0$ and $\bar{s} = 0$). $\bar{R} = 1$ is satisfied if and only if the circular data have a high concentration at a specific point ($\bar{c} + \bar{s} = 1$).

The concentration parameter k of the circular observations is estimated according to the following formula (Fisher, 1993):

$$\hat{k}=\begin{cases}2\bar{R}+\bar{R}^3+\frac{5}{6}\bar{R}^5 & \text{if } \bar{R}<0.53\\ -0.4+1.39\bar{R}+\frac{0.43}{\left(1-\bar{R}\right)} & \text{if } 0.53\le\bar{R}<0.85\\ \left(\bar{R}^3-4\bar{R}^2+3\bar{R}\right)^{-1} & \text{if } \bar{R}\ge 0.85\end{cases} \tag{4}$$

Methods

Detection of a Single Outlier by Existing Methods

The following methods for the detection of single outlier will be compared in univariate circular data with the suggested procedure.

Mardia Statistic

Mardia (1975) suggested a statistic to identify a single outlier in univariate circular data. The outlier was considered to be the observation that is the most influential on the resultant length. Therefore,

$$Mar = \min\left\{\frac{n-1-R_{(-i)}}{n-R}\right\} \tag{5}$$

where $R_{(-i)}$ is the resultant length after omitting the i^{th} observation and R is the resultant length for the full data set.

M Statistic

Collett (1980) suggested the M statistic to detect an outlier in univariate circular data. This is given as

$$M = \max\left\{\frac{R_{(-i)}-R+1}{n-R}\right\} = \frac{R_k - R+1}{n-R} \tag{6}$$

where $R_k = \max\{R_{(-i)}\}$.

A Statistic

Abuzaid et al. (2009) used the circular distance between the circular observations ϑ_i and ϑ_j as suggested by Rao (1969). This is given as

$$d_{ij} = 1-\cos\left(\vartheta_i - \vartheta_j\right)$$

where $d_{ij} \in [0, 2]$. The sum of all circular distances of the ϑ_j to all other observations is given by

$$D_j = \sum_{i=1}^{n} \left(1 - \cos\left(\vartheta_i - \vartheta_j\right)\right), \quad j = 1, 2, \ldots, n$$

Abuzaid et al. (2009) argued that if the observation θ_j is an outlier (and so lies far away from the other observations), the value of D_j will increase. Therefore, the A statistic to detect an outlier in the circular univariate data is based on the average circular distance

$$\frac{D_j}{n-1}$$

when omitting the observation θ_j. The statistic is given as follows:

$$A = \max\left\{\frac{D_j}{2(n-1)}\right\}, \quad j = 1, 2, \ldots, n \tag{7}$$

where $A \in [0, 1]$. Jammalamadaka and SenGupta (2001) considered the circular distance between any two circular data points to be the smallest arc between them on the circumference. They calculate the circular distance between ϑ_i and ϑ_j as follows:

$$d_{ij}^{*} = \pi - \left|\pi - \left|\vartheta_i - \vartheta_j\right|\right| \tag{8}$$

where $d_{ij}^{*} \in [0, \pi]$.

Abuzaid (2010) used equation (8) to calculate the sum of all circular distances from the observation ϑ_j to all other observations. This is given as

$$D_j^{*} = \sum_{i=1}^{n} \left(\pi - \left|\pi - \left|\vartheta_i - \vartheta_j\right|\right|\right), \quad j = 1, 2, \ldots, n$$

An alternative statistic is given by

$$A^{*} = \max\left\{\frac{D_j^{*}}{n-1}\right\}, \quad j = 1, 2, \ldots, n$$

where $A^{*} \in [0, \pi]$. Abuzaid expected that the statistic A^{*} has a similar performance to the A statistic.

Chord Statistic

Abuzaid (2010) used the geometrical properties of the chord of a circle to develop an alternative test to identify an outlier in circular univariate data. A chord is a segment that connects two different points on circumference of the circle. The length of the chord between ϑ_i and ϑ_j can be calculated using the following formula:

$$\operatorname{crd}\left(d_{ij}\right) = 2r\sin\frac{d_{ij}^{*}}{2r}$$

where r is the radius, so in the unit circle $r = 1$, and d_{ij}^{*} is the smallest arc length between ϑ_i and ϑ_j, which is calculated from equation (8).

In the unit circle, calculating S_j, the sum of the lengths of all the chords passing through observation ϑ_j, was proposed as

$$S_j = \sum_{i=1}^{n}\operatorname{crd}\left(d_{ij}\right) = 2\sum_{i=1}^{n}\sin\frac{d_{ij}^{*}}{2}, \quad j = 1, 2, \ldots, n$$

where $0 \leq S_j \leq 2(n - 1)$. When $d_{ij}^{*} = 0$,

$$\sum_{i=1}^{n}\sin\frac{d_{ij}^{*}}{2} = 0$$

while

$$\sum_{i=1}^{n}\sin\frac{d_{ij}^{*}}{2} = n-1$$

when $d_{ij}^{*}=\pi$. Therefore, if S_j has the maximum value, this suggests that ϑ_j is a candidate for the outlier. The chord statistic is given by

$$Chord=\max\left\{\frac{S_j}{2(n-1)}\right\},\quad j=1,2,\ldots,n \tag{9}$$

The Proposed Method

A robust circular distance *RCDu* statistic is now proposed to identify outliers in the circular data. It depends on two main points: first, the fact that the outliers in circular data may not be extreme values and second, an important property of the von Mises distribution is to be symmetric about the mean direction. However, the circular median is more efficient than the mean direction when the circular data have outliers (Ducharme & Milasevic, 1987). He and Simpson (1992) recommended the circular median is more robust than mean direction when the data do not follow von Mises distribution. Therefore, use the circular distance between any observation and circular median as a statistic to detect single and multi-outliers.

Suppose ϑ_1, ϑ_2,…, ϑ_n are circular observations located on the circumference of a unit circle. To apply the proposed procedure, there are several possible ways to calculate the circular distance $dist_{(i)}$ between ϑ_i and the circular median *med* because of the circular geometry of the data. The cases are as follows:

i. If $0 \le med \le \pi$

$$dist_{(i)}=\begin{cases}\left|\vartheta_i-med\right| & \text{if } \left|\vartheta_i-med\right|\le\pi\\ 2\pi-\vartheta_i+med & \text{if } \left|\vartheta_i-med\right|>\pi\end{cases} \tag{10}$$

ii. If $\pi \le med \le 2\pi$

$$dist_{(i)}=\begin{cases}\left|\vartheta_i-med\right| & \text{if } \left|\vartheta_i-med\right|\le\pi\\ 2\pi-med+\vartheta_i & \text{if } \left|\vartheta_i-med\right|>\pi\end{cases} \tag{11}$$

If ϑ_i is an outlier then $dist_{(i)}$ is expected to be relatively large. Therefore, the cut-off point is given by

$$RCDu=\max\left(dist\right) \tag{12}$$

Table 1. Cut-off points for *RCDu* statistic with 10%

n	*k* = 2	3	4	5	6	7	8	9	10	12	15	20
10	2.550	1.900	1.490	1.280	1.130	1.040	0.974	0.910	0.848	0.775	0.710	0.598
20	2.800	2.190	1.660	1.440	1.270	1.150	1.070	1.000	0.938	0.843	0.756	0.646
30	2.820	2.350	1.760	1.520	1.330	1.780	1.090	1.010	0.974	0.860	0.762	0.662
40	2.850	2.500	1.860	1.560	1.350	1.240	1.130	1.060	1.020	0.911	0.797	0.786
50	3.000	2.610	1.930	1.610	1.400	1.270	1.170	1.090	1.040	0.925	0.818	0.698
60	3.030	2.660	1.990	1.630	1.420	1.300	1.190	1.110	1.080	0.940	0.835	0.715
70	3.060	2.700	2.030	1.660	1.480	1.320	1.210	1.130	1.060	0.960	0.849	0.725
80	3.060	2.750	2.090	1.720	1.500	1.340	1.230	1.140	1.080	0.981	0.858	0.733
90	3.080	2.800	2.120	1.750	1.510	1.350	1.250	1.150	1.100	0.990	0.866	0.740
100	3.080	2.830	2.150	1.750	1.520	1.360	1.260	1.170	1.110	0.995	0.872	0.752
110	3.090	2.860	2.190	1.770	1.550	1.380	1.270	1.180	1.120	0.996	0.886	0.755
120	3.090	2.880	2.230	1.790	1.560	1.390	1.280	1.190	1.130	1.010	0.895	0.759
130	3.100	2.890	2.270	1.810	1.570	1.390	1.300	1.200	1.140	1.030	0.897	0.770
140	3.100	2.900	2.270	1.850	1.580	1.410	1.300	1.210	1.140	1.010	0.905	0.776
150	3.110	2.930	2.310	1.840	1.590	1.420	1.320	1.220	1.150	1.040	0.917	0.782
160	3.110	2.940	2.330	1.870	1.610	1.440	1.320	1.230	1.150	1.040	0.913	0.784
170	3.110	2.960	2.380	1.860	1.620	1.440	1.330	1.240	1.160	1.040	0.917	0.788
180	3.110	2.970	2.420	1.880	1.630	1.450	1.340	1.250	1.170	1.050	0.931	0.792
190	3.110	2.980	2.390	1.920	1.630	1.460	1.350	1.250	1.170	1.060	0.928	0.795
200	3.110	2.990	2.450	1.910	1.640	1.480	1.350	1.260	1.170	1.070	0.932	0.797

Consequently, ϑ_i is identified as an outlier if $dist_{(i)}$ exceeds the cut-off point. We will depend on a triple measure of robustness to evaluate the proposed method:

i. Proportion of outliers detected.
ii. Rate of masking.
iii. Rate of swamping.

This triple measure of robustness is very popular in the robustness literature for evaluating a particular method. A high proportion of outliers detected, and low masking and swamping rates, are always considered to be good robustness properties for any outlier detection statistic.

Results

The Cut-Off Point for the *RCDu* Statistic

The *RCDu* statistic has no simple known distributional form. Therefore, a series of simulation studies of univariate circular data are carried out to find the cut-off point for the *RCDu* statistic using Monte Carlo methods. The same procedure has been used by Jammalamadaka and SenGupta (2001) and Abuzaid et al. (2009). Twenty different sample sizes of n = 10, 20, 30,…, 200 and twelve values of concentration parameter k = 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 15, 20 are used in these simulation studies. First, generate a set of circular data such that $\vartheta \sim$ vM(0, k), for each sample size n and concentration parameter k. Then the *RCDu* statistic is calculated. The process is replicated 5000 times to generate the *RCDu* statistic for each combination of sample size n and concentration parameter k. Finally, the 10% and 5% upper points of *RCDu* are tabulated in Tables 1 and 2, respectively. These tabulated values for different sample sizes and concentrations can be used as cut-off points for the proposed statistic. However, it is possible to find the cut-off points for any sample size and concentration parameter. R codes to generate any cut-off points are available from the corresponding author.

Table 2. Cut-off points for *RCDu* statistic with 5%

n	***k* = 2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**12**	**15**	**20**
10	2.760	2.200	1.650	1.430	1.160	1.100	1.080	1.020	0.948	0.857	0.770	0.663
20	2.960	2.490	1.850	1.600	1.400	1.250	1.170	1.090	1.020	0.913	0.809	0.701
30	2.980	2.650	1.980	1.670	1.460	1.290	1.180	1.100	1.070	0.934	0.827	0.712
40	3.000	2.760	2.100	1.720	1.460	1.350	1.230	1.140	1.100	0.968	0.984	0.838
50	3.060	2.840	2.170	1.780	1.520	1.380	1.270	1.170	1.110	0.982	0.865	0.744
60	3.080	2.880	2.230	1.780	1.540	1.410	1.270	1.190	1.130	1.010	0.886	0.767
70	3.100	2.920	2.280	1.840	1.600	1.430	1.300	1.210	1.140	1.020	0.901	0.783
80	3.100	2.950	2.390	1.900	1.630	1.450	1.320	1.220	1.160	1.040	0.917	0.776
90	3.110	2.960	2.440	1.930	1.630	1.460	1.340	1.230	1.170	1.040	0.925	0.790
100	3.110	2.990	2.510	1.930	1.640	1.470	1.360	1.240	1.190	1.050	0.933	0.806
110	3.120	3.000	2.440	1.950	1.680	1.480	1.370	1.270	1.190	1.060	0.942	0.797
120	3.120	3.010	2.490	1.970	1.680	1.500	1.380	1.280	1.200	1.090	0.950	0.807
130	3.120	3.020	2.550	1.990	1.690	1.510	1.390	1.290	1.220	1.090	0.952	0.814
140	3.120	3.030	2.540	2.010	1.710	1.540	1.390	1.290	1.210	1.080	0.958	0.818
150	3.130	3.020	2.610	2.030	1.720	1.530	1.410	1.300	1.210	1.090	0.963	0.831
160	3.120	3.040	2.610	2.060	1.740	1.540	1.400	1.310	1.220	1.120	0.970	0.834
170	3.120	3.040	2.660	2.040	1.740	1.560	1.410	1.320	1.230	1.110	0.974	0.838
180	3.130	3.060	2.730	2.060	1.750	1.560	1.430	1.330	1.250	1.120	0.987	0.843
190	3.120	3.060	2.660	2.130	1.750	1.560	1.440	1.330	1.250	1.120	0.986	0.841
200	3.120	3.060	2.740	2.090	1.760	1.570	1.450	1.340	1.250	1.120	0.990	0.845

The Performance of the *RCDu* Statistic

For a Single Outlier

The performance of the *RCDu* statistic compared with existing statistics, *Mardia*, *M*, *A*, and *Chord*, are compared using Monte Carlo simulations. The study parameters were sample sizes $n = 20$, 60, 100, and 150 with six concentration parameters, $k = 2$, 3, 5, 6, 8, and 10. The data are contaminated with a single outlier (ϑc) defined by

$$\vartheta \mathrm{c} = \vartheta + \lambda \pi \operatorname{mod}\left(2\pi\right) \tag{13}$$

where λ is the degree of contamination, ($0 \leq \lambda \leq 1$).

If $\lambda = 0$, there is no contamination at position [*d*]. If $\lambda = 1$, the circular observation is located at the anti-mode of its initial location.

Replicate these processes 3000 times for all combinations of the sample size and concentration parameter with $\lambda = 0.8$. Figure 1 gives the proportions of outliers detected and the rates of masking and swamping for the 10% and 5% of cut-off points for the sample sizes $n = 60$.

It can be seen that, for small values of concentration parameters, the performance of all the methods is relatively low. This is because the circular data will be more spread around the circumference of the circle for low values of the concentration parameter. Therefore, it is very difficult to detect outliers in this case (Collett, 1980). The proportions of outliers detected for the *A* and *Chord* statistics are close to those for the proposed *RCDu* statistic and have the highest proportions of detection outliers. Consequently, the *RCDu*, *A*, and *Chord* statistics have the lowest rates of masking. There is not swamping for all combinations with 10% and 5% cut-off points because the rates of swamping are equal to 0. The results for $n = 20$, 100, and 150 were consistent with the results in Figure 1, so are not shown. Interested readers can request the corresponding author to provide more results.

For Multi-Outliers

In order to test performance of our statistic for different ratios of contamination, three ratios of contamination (5%, 10%, and 20%) were select, replicated 3000 times for the same combinations of sample sizes and concentration parameters with 10% and 5% of cut-off points of the *RCDu* statistic. The results of the triple measure of $n = 60$ are given in Figure 2. The performance of the *RCDu* statistic is relatively low for small values of concentration parameter (for the same reason that we mentioned above). However, the *RCDu* statistic successfully identifies outliers for $k \geq 5$ for different ratios of contamination. It has the highest proportion of detection of outliers and the lowest rate of masking. There is no swamping for all ratios of contamination for all combinations. The same results for n = 20, 100, and 150 were obtained (also available from the corresponding author).

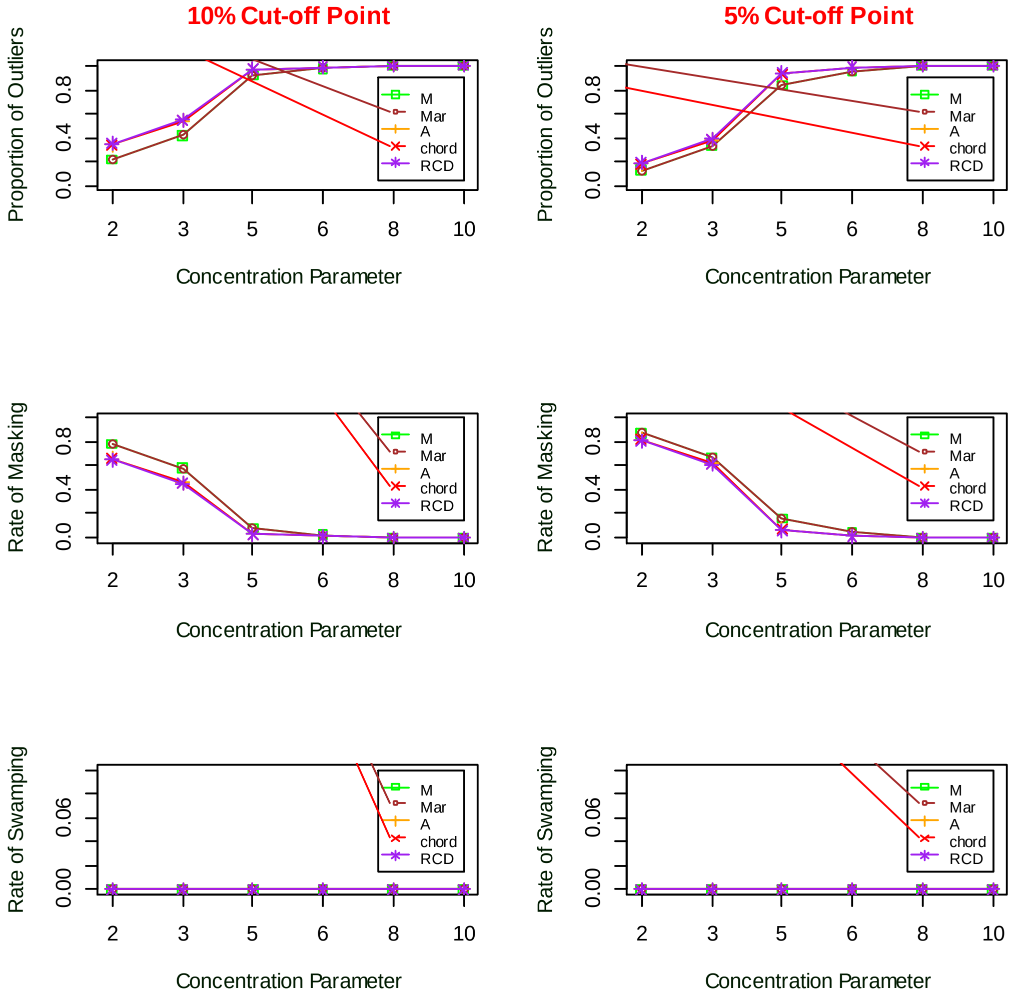

Figure 1. The proportion of a single outlier detected, and rate of masking and swamping, for different statistics with 10% and 5% of cut-off points for n = 60

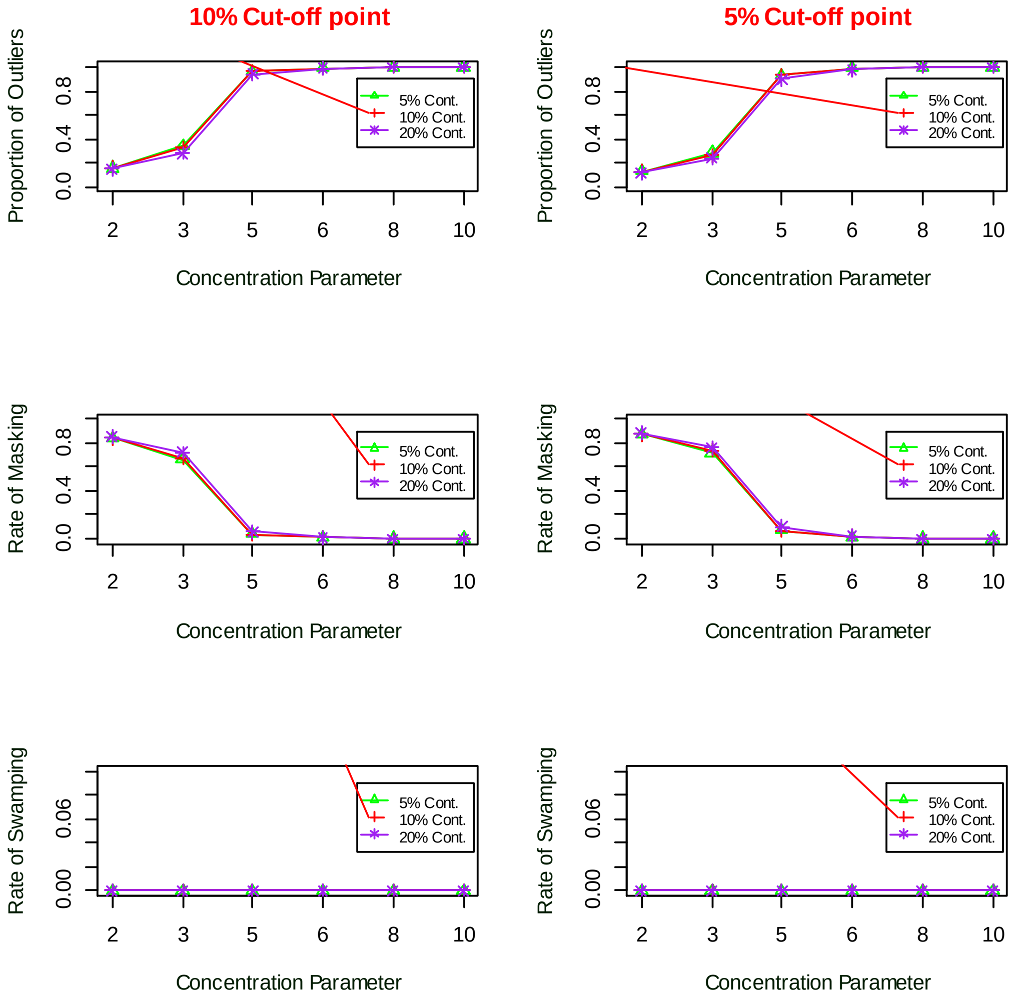

Figure 2. The proportion of outliers detected, and rates of masking and swamping, for 5%, 10%, and 20% of contamination with 10% and 5% of cut-off points for n = 60

Illustrative Examples

Example 1: A sample of 14 frogs was collected from the mud flats near Indianola, Mississippi. After 30 hours, the frogs were released and their directions were taken. Abuzaid et al. (2009) tested these circular data and detected that the observation numbered 14 is an outlier. The original circular data were tested; then outliers were inserted to bring the ratio of contamination to 20% in order to test the performance of the statistic with a high ratio of contamination. The cutoff point is approximately 2.68 according to the results in Table 1 (with 10% cut off point and concentration parameter $\hat{k} = 2.18$). The proportion of outliers detected and rate of masking and swamping are tabulated in Table 3 for all methods.

The competing statistics failed to detect outliers with 20% of contamination. The $dist_{(i)}$ statistic is plotted in Figure 3 for the original data set and with contamination 20%, respectively. As noted in Figure 3a, observation 14 is classified as an outlier because $dist_{(14)} = 3.06$ exceeds the cut-off point. Also, in Figure 3b, note $dist_{(12)} = 2.87$, $dist_{(13)} = 2.73$, and $dist_{(14)} = 3.06$ exceed the cut-off point. Therefore, the observations numbered 12, 13, and 14 are classified as outliers.

Table 3. Comparison of the performance of the different statistics (frogs data)

		Mardia	***M***	***A***	***Chord***	***RCDu***
Original data set	Proportion of outlier	1	1	1	1	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 20%	Proportion of outlier	0	0	0	0	1
	Rate of masking	1	1	1	1	0
	Rate of swamping	0	0	0	0	0

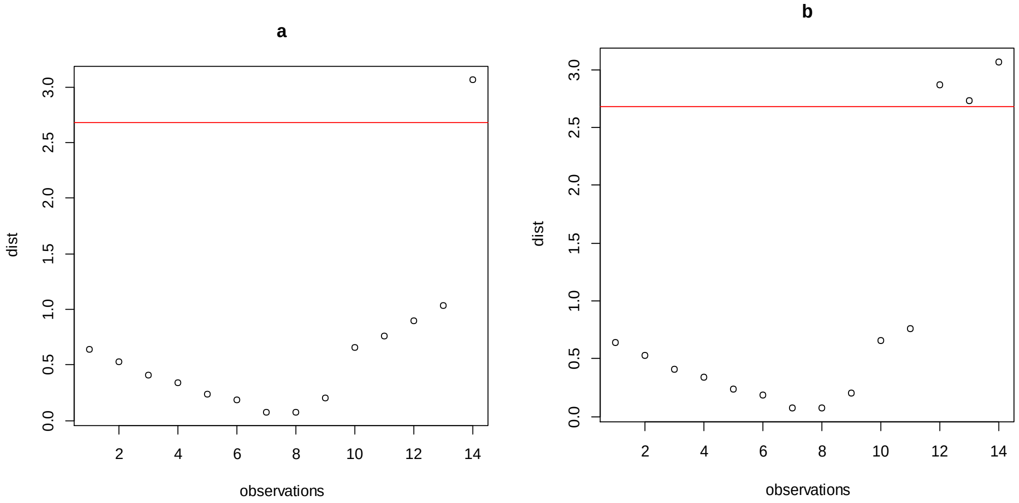

Figure 3. *dist*$_{(i)}$ statistic for frogs direction data

Table 4. Comparison of the performance of the different statistics (paleocurrent orientations data)

		Mardia	***M***	***A***	***Chord***	***RCDu***
Original data set	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 5%	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 10%	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0
Contamination 20%	Proportion of outlier	0	0	0	0	1
	Rate of masking	0	0	0	0	0
	Rate of swamping	0	0	0	0	0

Example 2: Next, consider the data given by Jammalamadaka and SenGupta (2001, p. 238). The sample size (40) represent measurements of the first sample of paleocurrent orientations from three bedded sandstone layers, measured on the Belford Anticline, New South Wales. They detected that observation number 24 is an outlier. In order to test performance of the statistics with different ratios of contamination, we insert outliers to bring the ratio of contamination to 5%, 10%, and 20%. The 10% cut-off point with concentration parameter $\hat{k} = 2$ is 2.85. The proportion of detection of outliers and rate of masking and swamping are tabulated in Table 4.

The other statistics fail to detect outliers with all ratios of contamination. However, the *RCDu* statistic succeeds in identifying all of them without any swamping. The $dist_{(i)}$ is plotted in Figure 4 for the original data set and with 5%, 10%, and 20% contamination.

Note in Figure 4a that $dist_{(24)} = 2.854$ exceeds the cut-off point. Therefore, it is classified as an outlier. This identification coincides with the identification by Jammalamadaka and SenGupta (2001) that this data point is an outlier. In Figure 4b, note that $dist_{(24)} = 2.915$ and $dist_{(40)} = 3.037$ exceed the cut-off point. Therefore, the observations numbered 24 and 40 are classified as outliers. Also, as expected, in Figure 4c, $dist_{(1)} = 3.124$, $dist_{(17)} = 3.072$, $dist_{(24)} = 3.054$, and $dist_{(40)} = 2.932$ exceed the cut-off point. Therefore, they are classified as outliers. The results $dist_{(1)} = 3.089$, $dist_{(9)} = 3.037$, $dist_{(10)} = 2.967$, $dist_{(17)} = 3.142$, $dist_{(24)} = 2.985$, $dist_{(33)} = 3.142$, $dist_{(39)} = 3.124$, and $dist_{(40)} = 2.862$ are greater than the cut-off point, so they are detected as outliers. They are given in Figure 4d.

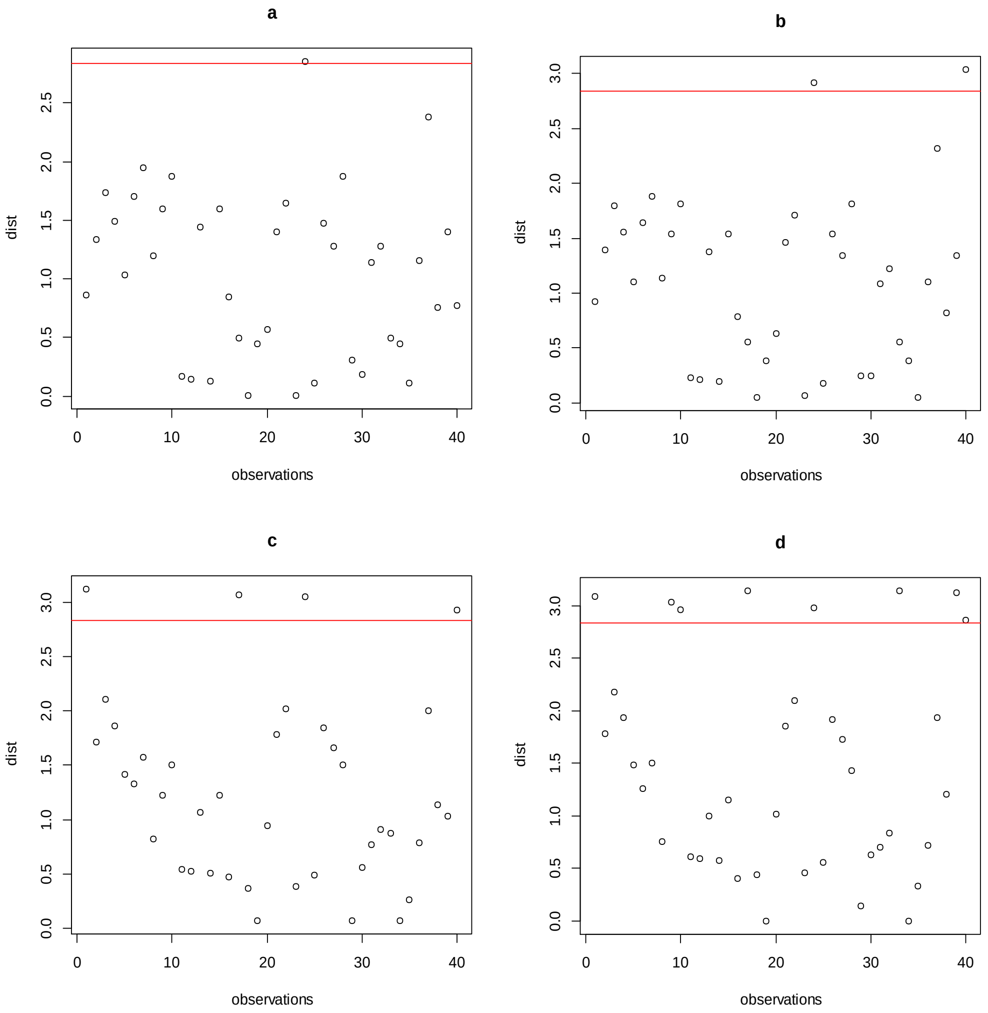

Figure 4. *dist*$_{(i)}$ statistic for paleocurrent orientations data

Example 3: In the final example, consider the wind direction data with sample size 129 that is considered by Abuzaid (2010, pp. 152-153). The data measurements were recorded over a period of 22.7 days along the Holderness coastline (the North Sea coast of Humberside in the United Kingdom). We insert outliers to bring the ratio of contamination to 5%, 10%, and 20%. The estimated concentration parameter is $\hat{k} = 7$, so the 10% cut-off point is equal to 1.39. The proportion of outliers detected and the rates of masking and swamping are tabulated in Table 5 for all the methods.

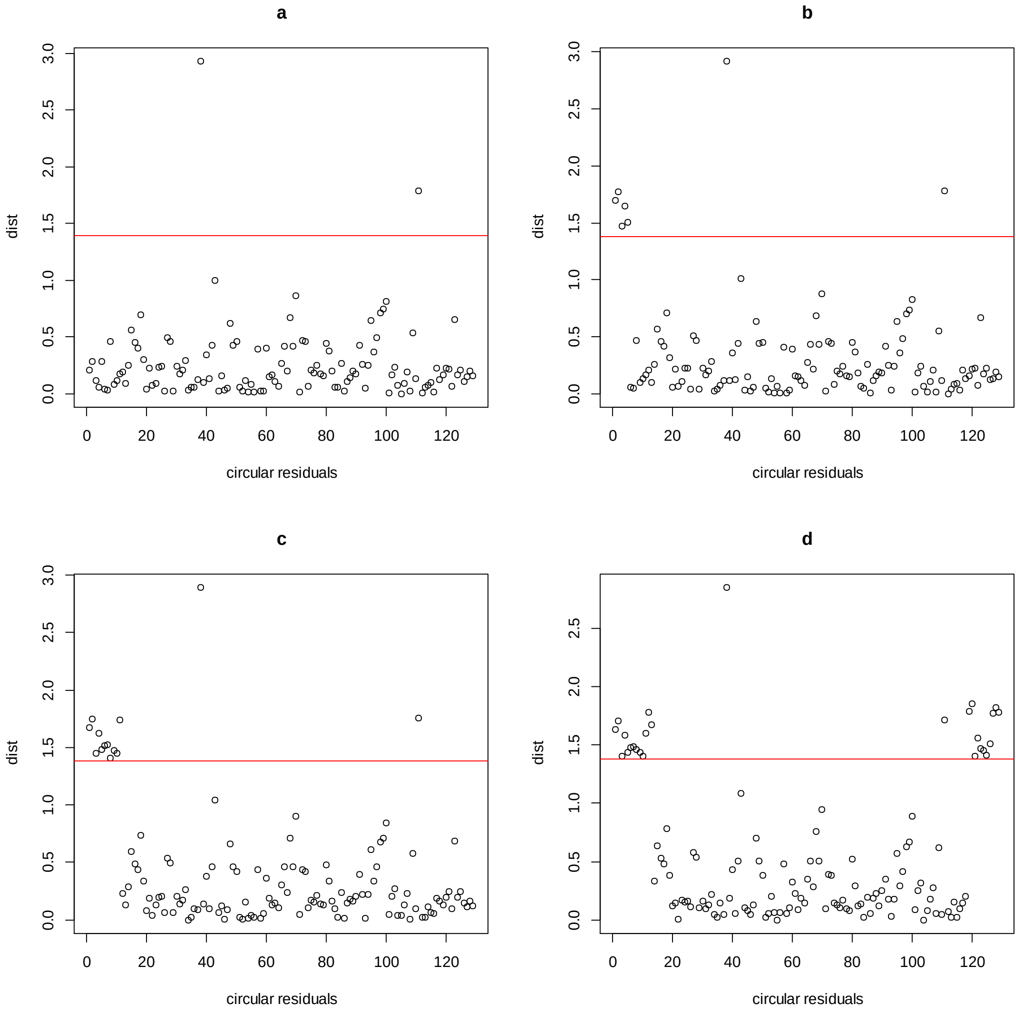

Figure 5. *dist$_{(i)}$* statistic for circular residuals

To illustrate this, the *dist$_{(i)}$* statistic is plotted in Figure 5 for the original data set and with 5%, 10%, and 20% of contamination. All the statistics perform the same for the original data set. However, the *M* and *Mardia* statistics fail to detect outliers after contamination. Besides, the *A* and *Chord* statistics cannot identify all of the outliers at the 10% and 20% contamination levels. In contrast, the *RCDu* statistic has the greatest proportion of outliers detected without swamping for all ratios of contamination.

Table 5. Comparison of the performance of the different statistics (wind direction data)

		Mardia	***M***	***A***	***Chord***	***RCDu***
Original data set	Proportion of outlier	1.00	1.00	1.00	1.00	1.00
	Rate of masking	0.00	0.00	0.00	0.00	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00
Contamination 5%	Proportion of outlier	0.14	0.14	1.00	1.00	1.00
	Rate of masking	0.86	0.86	0.00	0.00	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00
Contamination 10%	Proportion of outlier	0.08	0.08	0.92	0.77	1.00
	Rate of masking	0.92	0.92	0.08	0.23	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00
Contamination 20%	Proportion of outlier	0.00	0.00	0.73	0.46	1.00
	Rate of masking	1.00	1.00	0.27	0.54	0.00
	Rate of swamping	0.00	0.00	0.00	0.00	0.00

Conclusion

The *RCDu* statistic was proposed to detect single and multi-outliers in univariate circular data. The proposed statistic is evaluated based on the proportion of outliers detected and the masking and swamping rates. The proposed statistic has the highest proportion of outliers detected and the lowest rates of masking and swamping. Moreover, the proposed *RCDu* statistic is able to detect outliers in data with a high level of contamination. Also, the proposed statistic is successful in detecting outliers in a large data set. Hence, we suggest that the *RCDu* statistic should be used to detect outliers in univariate circular data.

References

Abuzaid, A. H. (2010). *Some problems of outliers in circular data* (Unpublished doctoral thesis). University of Malaya, Kuala Lumpur, Malaysia.

Abuzaid, A. H., Hussin, A. G., Rambli, A., & Mohamed, I. (2012). Statistics for a new test of discordance in circular data. *Communications in Statistics – Simulation and Computation, 41*(10), 1882-1890. doi: 10.1080/03610918.2011.624239

Abuzaid, A. H., Mohamed, I. B., & Hussin, A. G. (2009). A new test of discordancy in circular data. *Communications in Statistics – Simulation and Computation, 38*(4), 682-691. doi: 10.1080/03610910802627048

Barnett, V. & Lewis, T. (1978). *Outliers in statistical data*. New York, NY: Wiley.

Beckman, R. J., & Cook, R. D. (1983). Outlier..........s. *Technometrics, 25*(2), 119-149. doi: 10.2307/1268541

Collett, D. (1980). Outliers in circular data. *Applied Statistics, 29*(1), 50-57. doi: 10.2307/2346410

Ducharme, G. R., & Milasevic, P. (1987). Some asymptotic properties of the circular median. *Communications in Statistics – Theory and Methods*, *16*(3), 659-664. doi: 10.1080/03610928708829394

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge, UK: Cambridge University Press. doi: 10.1017/cbo9780511564345

Fisher, N. I., Lewis, T., & Willcox, M. E. (1981). Tests of discordancy for samples from Fisher's distribution on the sphere. *Applied Statistics, 30*(3), 230-237. doi: 10.2307/2346346

Hamelryck, T., Mardia, K., & Ferkinghoff-Borg, J. (2012). *Bayesian methods in structural bioinformatics*. Berlin, Germany: Springer-Verlag. doi: 10.1007/978-3-642-27225-7

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J., & Stahel, W. (1986). Robust statistics: The approach based on influence functions. New York, NY: Wiley. doi: 10.1002/9781118186435

He, X., & Simpson, D. G. (1992). Robust direction estimation. *The Annals of Statistics*, *20*(1), 351-369. doi: 10.1214/aos/1176348526

Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in circular statistics*. Singapore: World Scientific Publishing. doi: 10.1142/4031

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society*. *Series B (Methodological), 37*(3), 349-393. Available from http://www.jstor.org/stable/2984782

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. Chichester, UK: John Wiley & Sons Ltd. doi: 10.1002/9780470316979

Mohamed, I. B., Rambli, A., Khaliddin, N., & Ibrahim, A. I. N. (2015). A new discordancy test in circular data using spacings theory. *Communications in Statistics – Simulation and Computation, 45*(8), 2904-2916. doi: 10.1080/03610918.2014.932799

Otieno, B. S., & Anderson-Cook, C. M. (2003). *Hodges-Lehmann estimator of preferred direction for circular* (Technical Report 03-3). Blacksburg, VA: Virginia Tech Department of Statistics.

Otieno, B. S., & Anderson-Cook, C. M. (2005). Effect of position of an outlier on the influence curve of the measures of preferred direction for circular data. *Journal of Modern Applied Statistical Methods*, *4*(1), 81-89. doi: 10.22237/jmasm/1114906140

Rao, J. S. (1969). *Some contributions to the analysis of circular data* (Unpublished doctoral thesis). Calcutta, India: Indian Statistical Institute.