Pertanika J. Sci. & Technol. 25 (4): 1147 - 1158 (2017)

SCIENCE & TECHNOLOGY

Journal homepage: http://www.pertanika.upm.edu.my/

Adjusting Outliers in Univariate Circular Data

Mahmood, Ehab A.[1], Rana, Sohel[1]*, Hussin, Abdul Ghapor[2] and Midi, Habshah[1]

[1]*Department of Mathematics, Institute for Mathematical Research, Universiti Putra Malaysia, 43400 UPM, Serdang, Selangor, Malaysia*
[2]*Faculty of Defence Science and Technology, National Defence University of Malaysia, 57000 UPNM, Kuala Lumpur, Malaysia*

ABSTRACT

Circular data analysis is a particular branch of statistics that sits somewhere between the analysis of linear data and the analysis of spherical data. Circular data are used in many scientific fields. The efficiency of the statistical methods that are applied depends on the accuracy of the data in the study. However, circular data may have outliers that cannot be deleted. If this is the case, we have two ways to avoid the effect of outliers. First, we can apply robust methods for statistical estimations. Second, we can adjust the outliers using the other clean data points in the dataset. In this paper, we focus on adjusting outliers in circular data using the circular distance between the circular data points and the circular mean direction. The proposed procedure is tested by applying it to a simulation study and to real data sets. The results show that the proposed procedure can adjust outliers according to the measures used in the paper.

Keywords: Directional data, circular data, circular mean, circular distance, outlier

INTRODUCTION

In many different scientific fields such as geology, biology, meteorology, physics, psychology, image analysis and medicine, measurements are directions. For instance, in geology, a researcher may be interested in the direction of the earth's magnetic pole, while in biology, measurements may be made of the direction of birds' migration or the orientation of animals. A set of these directions is defined as *directional data*. There are two types of directional data. First, data may be represented in two dimensions on the circumference of a unit circle; this type of data is called *circular data*. Circular data can be represented as clockwise or anti-clockwise measurements and may be measured either in degrees,

ARTICLE INFO

Article history:
Received: 14 April 2016
Accepted: 13 April 2017

E-mail addresses:
eee.mahmood@gmail.com (Mahmood, Ehab A.),
sohel_rana@upm.edu.my (Rana, Sohel),
abdulghapor@gmail.com (Hussin, Abdul Ghapor),
habshahmidi@gmail.com (Midi, Habshah)
*Corresponding Author

ISSN: 0128-7680

distributed in the interval [0°-360°), or in radians, in the interval [0-2π). Second, the data may be represented in three dimensions by two angles as points on the surface of a unit sphere or as points on the earth's surface, measured according to longitude and latitude. These directional data are defined as *spherical data* (Jammalamadaka & SenGupta, 2001).

Statistical data may show some observations that are not consistent with other data, and these are defined as *outliers*. Researchers have demonstrated that the existence of outliers in statistical data causes some serious problems in statistical analysis. It has been confirmed that outliers cause misleading statistical results and estimations of parameters, and may not bring accurate predictions. In addition, the classical methods of statistical analysis consider some conditions, one of them that the statistical data should be free of outliers (Barnett & Lewis, 1978). Maximum Likelihood Estimator (MLE) is the most famous of classical methods to estimate model parameters for linear and circular data. However, we cannot use it to estimate parameters if the data have outliers. Researchers have suggested two main ways to deal with this problem. First, outliers are to be detected and either deleted or adjusted, after which classical statistical methods can be applied. Second, robust statistical methods can be applied. In linear data, outliers are observations that are extreme. By contrast, outliers in circular data may not have extreme values because the circular data are bound by parameters. Therefore, circular data have properties that are different from linear data (Lee, 2010). For example, if we have two circular data points at 50° and 310°, then the arithmetic mean using the linear measure is equal to 180°. Nonetheless, the circular mean direction is equal to 0° using the geometrical theory of the circle. In addition, the smallest circular data point coincides with the largest point i.e. 0=2π and the measurement is periodic, with ϑ being the same as ϑ+p*2π for any integer *p*. The statistical measures that we apply to linear data cannot be used for circular data because of these geometrical properties of circular data.

The literature indicates that few researchers are aware of how to detect outliers in circular data. Furthermore, to date no one has suggested a procedure for adjusting outliers in circular 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. He found that for small samples sizes it is better to use the *C* and *D* statistics. Bagchi and Guttman (1990) used a Bayesian approach to identify outliers in circular data and to estimate the mean direction and the concentration parameters where the circular data follow a von Mises distribution. Fisher (1993) summarised 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, for which 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, and, 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 et al., 1981). Their final factor was the exponential distribution. Some tests for this factor were suggested by Fisher et al. (1981).

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. These authors used *LMPI* for the circular data that were obtained by mixing a wrapped stable distribution with a circular uniform distribution, *WSM*. Second, they proposed using a likelihood ratio testing (*LRT*) approach to identify outliers in circular data. Abuzaid et al. (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 relied 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.

Abuzaid (2010) used the geometrical properties of the chord of a circle for detecting an outlier in univariate circular data. Rambli et al. (2012) adopted the *M, C, D* and *A* statistics to identify outliers in univariate circular data when the circular data follow a wrapped normal distribution. They found that the *A* statistic is the best for a large sample size. However, for a small sample size the *M* statistic is the best. Abuzaid (2012) analysed Mother's Day celebration data using circular statistics. He applied cluster analysis to the circular data to define possible clusters and to detect outliers in univariate circular data. In addition, he used the *C* and *D* statistics as numerical statistics, as suggested by Collett (1980), and the boxplot as a graphical method to identify outliers. Abuzaid et al. (2012) 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. However, none of the previous researchers proposed how one could adjust the outliers.

In this paper, we propose a procedure to adjust outliers in univariate circular data to avoid their effects so that we can then apply classical circular statistical measures. This paper is arranged in the following sections. Section 2 describes the von Mises distribution and some important related formulas. Section 3 explains the proposed procedure for adjusting outliers. Section 4 illustrates the performance of our procedure. Section 5 gives an example based on real data to show how the proposed procedure can be used in real-life situations. Finally, in section 6, using all the numerical results, we conclude with the benefits of using the proposed procedure for univariate circular data.

THE VON MISES DISTRIBUTION AND ITS ESTIMATED PARAMETERS

The von Mises distribution is the most well-known of circular probability distributions. It is the same as normal distribution in linear data.

Let $\vartheta_1, \vartheta_2, \ldots\ldots, \vartheta_n$ be circular observations following a von Mises distribution with circular 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 et al. (2012):

$$g(\vartheta,\mu,k) = \frac{1}{2\pi I_0(k)} e^{k\cos(\vartheta-\mu)} \qquad [1]$$

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

$$I_0(k) = \frac{1}{2\pi}\int_0^{2\pi} e^{k\cos(\vartheta)}\, d\vartheta$$

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 (Mardia & Jupp, 2000), where:

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

The circular mean of the circular observations is estimated by maximum likelihood according to the following formula (Jammalamadaka & SenGupta, 2001):

$$\hat{\mu} = \begin{cases} tan^{-1}\left(\frac{s}{c}\right) & if\ \ c > 0,\ s \geq 0 \\ \frac{\pi}{2} & if\ \ c = 0,\ \ s > 0 \\ tan^{-1}\left(\frac{s}{c}\right) + \pi & if\ \ c < 0 \\ tan^{-1}\left(\frac{s}{c}\right) + 2\pi & if\ \ c \geq 0,\ \ s < 0 \\ undefined & if\ \ c = 0,\ \ s = 0 \end{cases} \qquad [2]$$

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

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:

$$\bar{R} = \sqrt{\bar{c}^2 + \bar{s}^2} \qquad [3]$$

where

$0 \leq \bar{R} \leq 1$

$\bar{c} = c/n$

$\bar{s} = s/n$

$\bar{R} = 0$ 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 maximum likelihood estimation of the concentration parameter k is given by the following formula (Fisher, 1993):

$$\hat{k} = \begin{cases} 2\bar{R} + \bar{R}^3 + \frac{5}{6}\bar{R}^5 & if\ \bar{R} < 0.53 \\ -0.4 + 1.39\bar{R} + \frac{0.43}{(1-\bar{R})} & if\ \ 0.53 \leq \bar{R} < 0.85 \\ (\bar{R}^3 - 4\bar{R}^2 + 3\bar{R})^{-1} & if\ \bar{R} \geq 0.85 \end{cases} \qquad [4]$$

PROPOSED PROCEDURE FOR ADJUSTING OUTLIERS

The circular distance between any two circular points is the smallest distance between them on the circumference of a circle and it lies in [0, π] (Jammalamadaka & SenGupta, 2001). In this paper, we assume that an outlier lies far from the circular mean. Therefore, we suggest to depend on the circular distance between circular data points and the circular mean as a procedure for adjusting.

Our proposed procedure is carried out in two stages. Let ϑ_1, ϑ_2, ….., ϑ_n be circular data with sample size *n*; in the first stage of our proposal, we adjust the circular distance between the outliers and the circular mean as follows:

i. Calculate the circular mean for clean data $\hat{\mu}_c$ (after delete outliers) in order to avoid their effects.

ii. Calculate the circular distance *dist* between the clean circular data points and $\hat{\mu}_c$ using the following formula:

If $\leq \hat{\mu}_c \leq \pi$:

$$dist_i \quad \begin{cases} |\vartheta_i - \hat{\mu}_c| & \text{if } |\vartheta_i - \hat{\mu}_c| \leq \pi \\ 2\pi - \vartheta_i + \hat{\mu}_c & \text{if } |\vartheta_i - \hat{\mu}_c| > \pi \end{cases} \qquad [5]$$

If $\pi < \hat{\mu}_c < 2\pi$:

$$dist_i = \begin{cases} |\vartheta_i - \hat{\mu}_c| & \text{if } |\vartheta_i - \hat{\mu}_c| \leq \pi \\ 2\pi - \hat{\mu}_c + \vartheta_i & \text{if } |\vartheta_i - \hat{\mu}_c| > \pi \end{cases} \qquad [6]$$

i = 1, 2, ….., n-out

where,

out : number of outliers

$0 \leq disti \leq \pi$

iii. Calculate mean of the circular distance MCD and max (*dist*). Clearly, max (*dist*) is not an outlier.

iv. Calculate the contaminated circular distance $dist_i$ (*cont*) between the outliers and $\hat{\mu}_c$.

v To adjust $dist_i$ (*cont*), we propose the following formula:

$$dist_i\,(adj) = (dist_i\,(cont) + MCD)\,/\,2 \qquad [7]$$

If $dist_i$ (*adj*) > max(*dist*), we continue to calculate a new $dist_i$ (*adj*) but we substitute $dist_i$ (*adj*) in place of $dist_i$ (*cont*) in Equation (7). We apply this progress until we have $dist_i$ (*adj*) ≤ max(*dist*). In this step, we try to minimise the value of *disti (cont)* to be less than max(*dist*) because max(dist) is not an outlier.

In the second stage, we depend on formulas (5) and (6) to calculate the adjusted value of outliers ϑi(adj) according to the following formula:

$$\vartheta_{\mathrm{i}}(\mathrm{adj}) = \hat{\mu}_c + dist_i\ (adj) \qquad [8]$$

In the adjustment procedure, we aim to minimise the circular distance between ϑ_{i} and the circular mean. Therefore, in Equation (8) we may subtract $dist_i$ (adj) from $\hat{\mu}_c$, instead of adding it, to minimise the circular distance. This depends on the position of the circular mean direction of the population on the circumference of the unit circle. For instance, if the circular mean for a particular population is equal to zero, we add $dist_i$ (adj) whenever $\hat{\mu}_c > \pi$, to make it closer to zero. Likewise, we subtract $dist_i$ (adj) whenever $\hat{\mu}_c < \pi$.

PERFORMANCE OF THE PROPOSED PROCEDURE

We examined the performance of our procedure by applying it to a series of simulation studies for univariate circular data using Monte Carlo methods. We depended on the four statistical measures to evaluate our suggestion: the bias of the circular mean, the bias of the concentration parameter *k*, mean resultant length $\overline{R}$ and mean of the circular distance *MCD*. The simulation studies were divided into four parts. First, we generated a set of circular data such that $\vartheta \sim vM(0,k)$ for three samples having the sizes 20, 40 and 60 and using six values of the concentration parameter $k = 0.5$, 1, 2, 3, 5 and 6. The statistical measures were calculated and called 'clean measures'. Second, the data were contaminated at position d using the following Equation:

$$\vartheta c_{[d]} = \vartheta_{[d]} + \lambda\pi \bmod(2\pi) \qquad [9]$$

where $\vartheta c_{[d]}$ is the contaminated circular observation at position [d], and λ is the degree of contamination, with $0 \leq \lambda \leq 1$.

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.

For all combinations of sample sizes and concentration parameters, we generated 5% and 10% of the contaminated data, with $\lambda = 0.8$. The statistical measures were calculated, called 'cont. measures'. Third, we deleted the *outliers* in the contaminated data to calculate the statistical measures, and these are called 'del. Measures'. Finally, we applied the proposed procedure and calculated the statistical measures, which are called 'adj. measures'. The process was replicated 5,000 times for each combination of sample size and concentration parameter *k*.

Figure 1 shows that the values of the bias of the estimated circular mean for the contaminated data are relatively large. In addition, the values of the bias of the estimated circular mean with *10%* contaminated data are larger than the values with *5%* for all combinations. In contrast, the values of the bias of the adjusted circular mean are relatively low and are close to the values of the biases for the data with no outliers and to the data with the outliers deleted. This was one of the measures used to evaluate our procedure.

We can notice in Figure 2 that there are no differences between the values of the biases of the estimated concentration parameters k, for $k \leq 3$. On the other hand, the contaminated data have large values of bias for $k > 3$. The values of the biases of the concentration parameter for the adjusted data are low and are close to the results with clean data and with the outliers deleted. This is more evidence of the success of our procedure for all combinations.

The results in Figure 3 show that there was a vast difference between the results of the contaminated data and those of the others especially with the increased ratio of contamination. In contrast, the results of the proposed procedure were close to 1 at high values of concentration parameters; they were as close as the clean data and the data with outliers deleted.

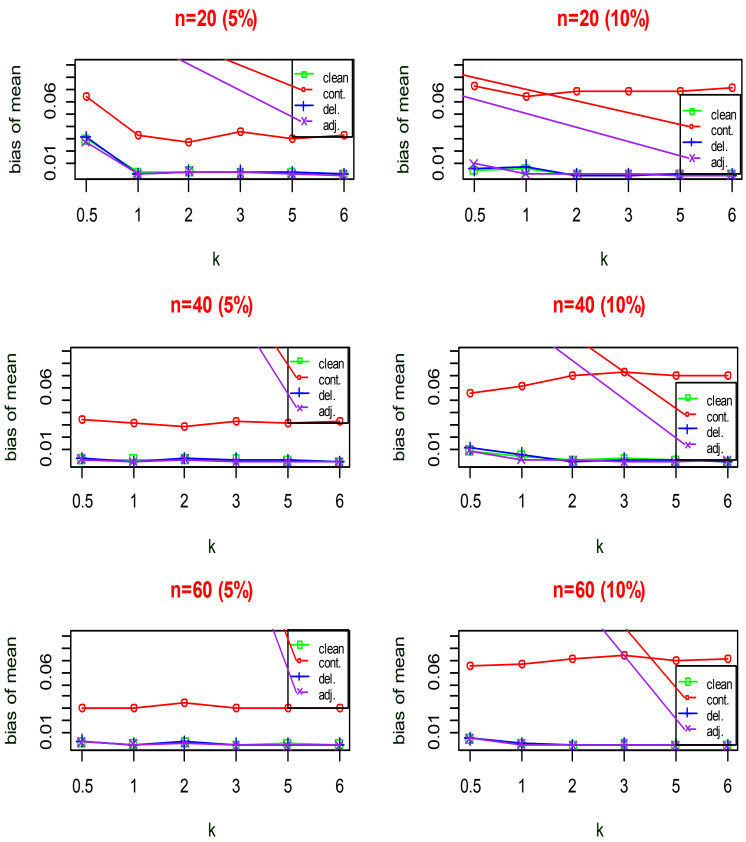

Figure 1. Bias of estimated circular mean

In Figure 4, the values of the *MCD* statistic increased for all combinations of contaminated data. Moreover, the values of the *MCD* statistic with *10%* outliers were larger than the values with *5%* outliers. However, our procedure can minimise these, giving values that are close for clean data and for data with the outliers deleted. This is the fourth measure used to test our procedure.

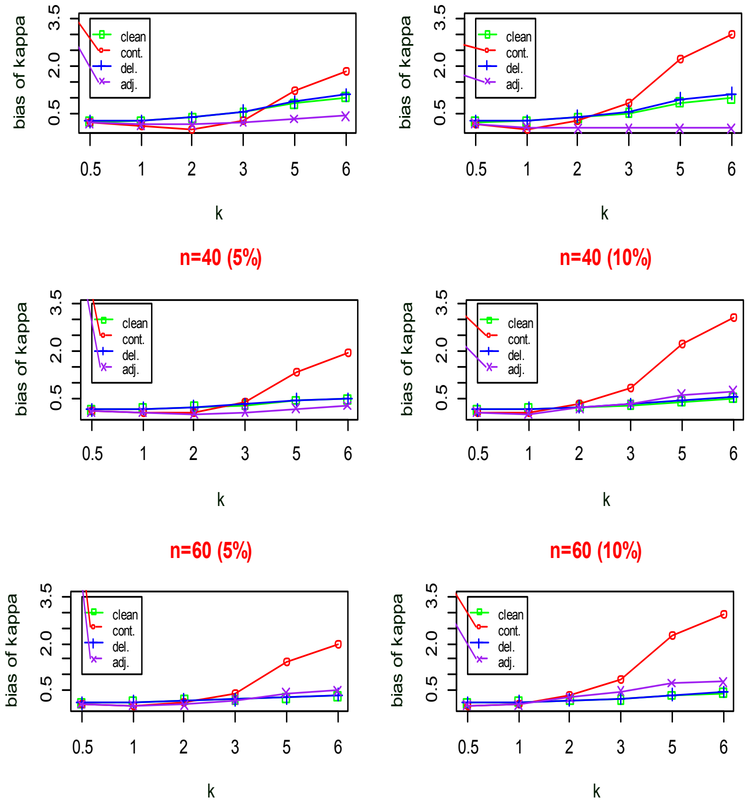

Figure 2. Bias of estimated kappa

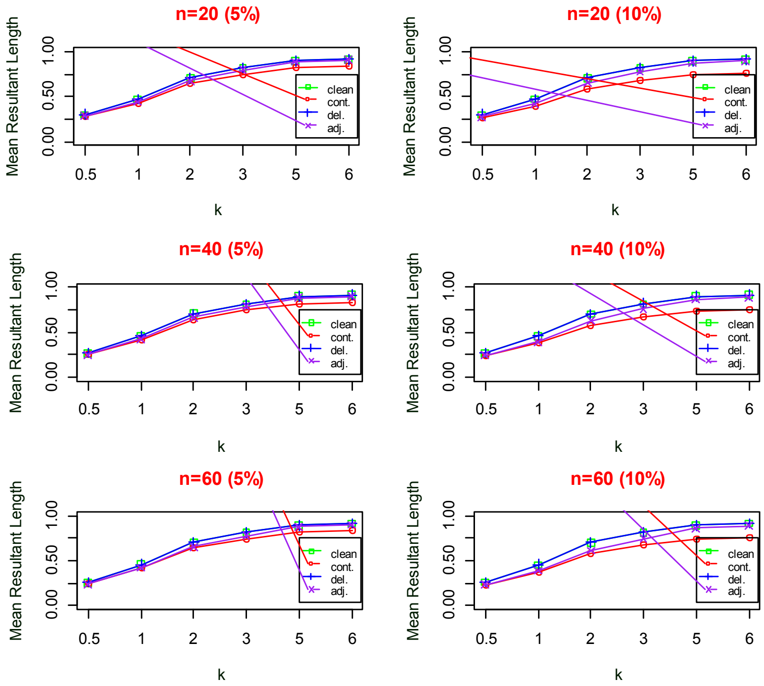

Figure 3. Mean resultant length

In summary, our procedure successfully adjusted the outliers for all combinations of the values of the concentration parameter *k* and all sizes of sample, with *5%* and *10%* contamination, according to the results for the statistical measures.

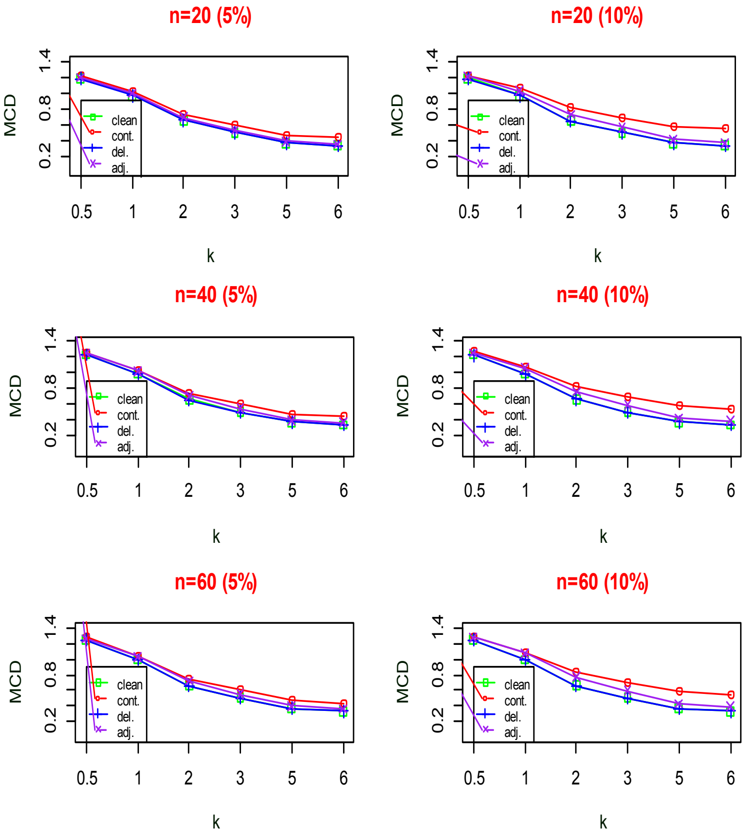

Figure 4. Mean circular distance

PRACTICAL EXAMPLE

We considered the sea stars data given by Fisher (1993). The data represent the measurements of the resultant directions of 22 sea stars for 11 days after they were displaced from their natural habitat. Fisher identified observation 13 as an outlier. To evaluate our proposed procedure, we applied the following steps. First, we estimated the circular mean $\hat{\mu}$, the concentration parameter k, the Mean Resultant Length $\bar{R}$ and mean of the circular distance *MCD* for three cases: with contaminated data; with the outliers deleted; and with the outliers adjusted. The results are shown in Table 1.

Table 1
Comparison of Measures for Three Cases (Sea Stars Data)

	cont.	del.	adj.
$\hat{\mu}$	0.054	0.023	6.27
$\hat{k}$	3.3	5.7	5.1
$\bar{R}$	0.83	0.91	0.90
MCD	0.43	0.33	0.35

The results shown in Table 1 show that the estimated circular mean after adjusting the outlier was closer to the circular mean with the outlier deleted than to the circular mean for the contaminated data. We could see a similar scenario for the concentration parameter *k*, Mean Resultant Length $\bar{R}$ and mean of the circular distance *MCD*. Second, we plotted the sea star data with the outlier and with the outlier adjusted. Figure 5 shows two cases for sea star data, with the outlier and with the outlier adjusted.

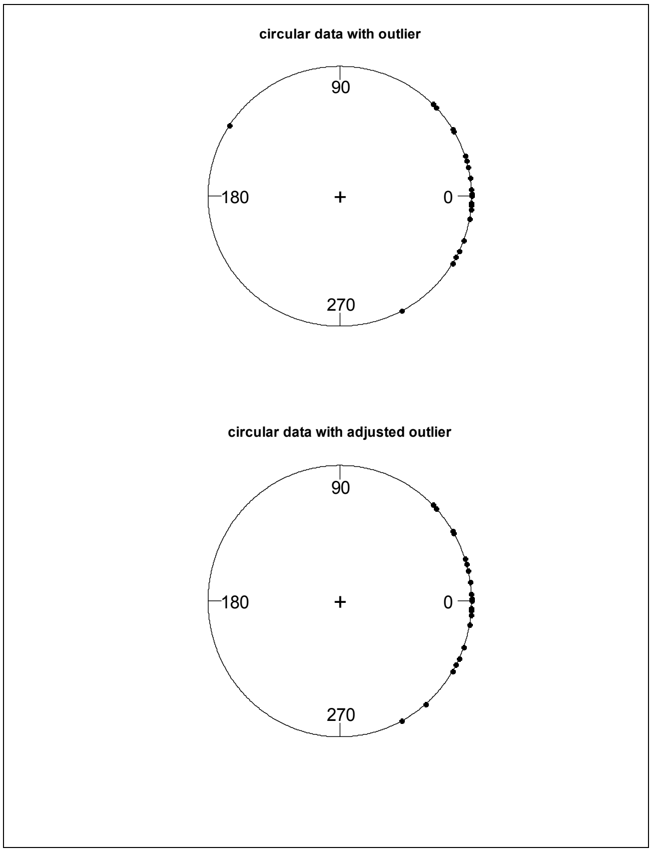

Figure 5. Sea star data with outlier and with adjusted outlier

It is clear that our procedure could significantly adjust the outlier and make the circular data more consistent. This is another piece of evidence that our procedure is successful in adjusting outliers.

CONCLUSION

In this paper, we proposed a new procedure for adjusting outliers in univariate circular data. In general, the proposed procedure performs well according to the statistical measures used in the paper. The proposed procedure decreases the bias of both the circular mean and the concentration parameter *k*. Moreover, the proposed procedure gives results of Mean Resultant Length as close as results of clean data and minimises *MCD* with low and high levels of contamination. The proposed procedure is also successful for different sample sizes. Hence, we suggest that our procedure should be used to adjust outliers in univariate circular data.

REFERENCES

Abuzaid, A. H. (2010). *Some problems of outliers in circular data* (Unpublished PhD thesis). Faculty of Science, University of Malaya, Malaysia.

Abuzaid, A. H. (2012). Analysis of Mother's Day celebration via circular statistics. *The Philippine Statistician, 61*(2), 39–52.

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.

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.

Bagchi, P., & Guttman, I. (1990). Spuriosity and outliers in directional data. *Journal of Applied Statistics 17*(3), 341–350.

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

Collett, D. (1980). Outliers in circular data. *Journal of Applied Statistics, 29*(1), 50–57.

Fisher, N. I. (1993). *Statistical analysis of circular data*. Cambridge: Cambridge University Press.

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

Hamelryck, T., Mardia, K., & Ferkinghoff-Borg, J. (2012). *Bayesian Methods in Structural Bioinformatics.* Berlin, Heidelberg: Springer-Verlag.

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

Lee, A. (2010). Circular Data. Interdisciplinary Reviews: *Computational Statistics, 2*(4), 477–486.

Mardia, K. V. (1975). Statistics of directional data. *Journal of the Royal Statistical Society, Series B, 37*(3), 349–393.

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

Rambli, A., Ibrahim, S., Abdullah, M. I., Hussin, A. G., & Mohamed I. (2012). On discordance test for the wrapped normal data. *Sains Malaysiana, 41*(6), 769–778.