# Outlier labeling via circular boxplot

Abuzaid, A. H.[1] Hussin, A. G.[2] Mohamed, I. B.[3]

[1]Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, MALAYSIA.
E-mail: alizaid@perdana.um.edu.my
[2]Center for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, MALAYSIA.
E-mail: ghapor@um.edu.my
[3]Institute of Mathematical Sciences, University of Malaya, 50603 Kuala Lumpur, MALAYSIA.
E-mail: imohamed@um.edu.my

**Keywords:** Boxplot, Circular data, Outliers, Overlapping.

**Abstract**

Boxplot is a simple and flexible graphical tool that has been widely used in exploratory data analysis. Its main application is to identify extreme values and outliers in linear univariate data sets. However, the standard boxplot for linear data set is not suitable to be used for circular data sets due to the bounded property of circular variables. In this paper, we propose and develop a boxplot for circular data sets based on five circular summary statistics which is called circular boxplot. In the process, several problems have been resolved. Firstly, we have overcome the problems of estimating the circular median, the first and second quartiles and overlapping areas between the upper and lower fences. Secondly, we resolve the problem of finding the appropriate boxplot criterion which is ($\nu IQR$=1.5$IQR$) in linear case, where $IQR$ is the interquartiles range and $\nu$ is the resistance constant. Through simulation studies, we identify the appropriate values of circular boxplot criterion which depends on the concentration parameter. The power of performances of the proposed boxplot is investigated. We then develop S-Plus subroutines to display the circular boxplot and apply the plot on a real circular data set.

## 1 Introduction

Visual display is an easy and informative technique to deal with the data. There are various graphs that enable statisticians to present their data such as histogram, pie chart, Q-Q plot and boxplot.

Boxplot is a simple and flexible graphical tool in exploratory data analysis developed by Tukey(1977). It consists of five-number summaries which are the smallest observation, lower quartile $Q_1$, median, upper quartile $Q_3$ and largest observation. Its main application is to identify extreme values and outliers in univariate data sets.

Most of text monographs employ 1.5$IQR$ boxplot criterion to detect outliers in data sets. In other words, any observation below $Q_1 - 1.5IQR$ or above $Q_3 + 1.5IQR$ are labeled as "outlier". Extensive research was conducted on the labeling of outlier by using boxplot. Hoaglin et al.(1986) investigated the performance of resistant rules for outlier labeling by using different values instead of $\nu = 1.5$. They concluded that the main resistant rule $\nu = 1.5$ has many advantages especially in avoiding the masking problems. Further, they considered the rule when $\nu = 3$ as extremely conservative. Ingelfinger et al.(1983) suggested the use of $\nu = 2$ while Sim et al.(2005) demonstrated that the use of resistant rule $\nu = 1.5$ or $\nu = 3$ is in general inappropriate for normal sample and is completely inappropriate for skewed distributions. So the value of $\nu$ differs from data set to another depending on the underlying distribution of the data set.

All known published works on the boxplot were conducted for linear variables. There is no relevant published work on circular variables, except Graedel(1977) who suggested the use of boxplot to improve the wind rose diagram.

There are many phenomena that can be considered as circular variables such as directions of wind and birds immigration. For such data, the linear techniques can no longer be used and special techniques are required as the identification of outliers in circular data requires different approach compared to the linear case. There are several methods available to detect outliers in circular data. Collett(1980) presented four different numerical tests of discordancy in circular data, namely $C, D, L$ statistics and improved $M$ statistics which was first proposed by Mardia(1975). Abuzaid et al.(2008) identified single

outlier in circular regression models based on the circular residuals by using different numerical and graphical methods.

In this paper the main objectives are to (i) construct a special boxplot for circular data called the circular boxplot (ii) use the circular boxplot in the detection of outliers in circular data and (iii) develop a subroutine in S-Plus environment to draw the circular boxplot.

In the next section we discusses the proposed construction of circular boxplot. Simulation and numerical studies are carried out in Sections 3 to estimate the appropriate values of resistance constant $\nu$. A numerical example is discussed in Section 4 to illustrate the application of circular boxplot in detecting outliers for circular regression based on circular residuals.

## 2 Summary statistics for constructing circular boxplot

Due to the unusual characteristics of circular variables, many relevant descriptive measures and display plots have been developed for example, circular histogram and stem-and-leaf diagrams. However, there is no known design of boxplot for circular variable. The difficulty of constructing the boxplot for circular variables arises from the complexity of determining the median. This is due to the bounded range of circular variables and the overlapping problem which occurs when the concentration parameter $\kappa$ of circular sample is small. These issues will be addressed in this section.

### 2.1 Median direction of circular variable

Fisher(1993) defined the median of circular variable as an axis which divides the data into two equal groups (median axis). Practically, the circular median is the observation $\phi$, which minimizes the summation of circular distances, $d(\phi) = \pi - \sum_{i=1}^{n} |\pi - |\theta_i - \phi||$. Consistently, Mardia and Jupp(2000) defined the median for a set of circular observations $\theta_1, \theta_2, ..., \theta_n$ as any point $\phi$, where half of the data lie in the arc of $[\phi, \phi + \pi)$ and the other points are nearer to $\phi$ than to $\phi + \pi$. In case of prior knowledge about the circular distribution, Mardia(1972) defined the median direction $\phi$ as the solution of

$$\int_{\phi}^{\phi+\pi} f(\theta)d\theta = \int_{\phi+\pi}^{\phi+2\pi} f(\theta)d\theta = 0.5.$$

### 2.2 Quartiles of circular variables

In order to construct boxplot for circular variable, we need to know the median, the first quartile $Q_1$ and the third quartile $Q_3$. Mardia(1972) defined the first and third quartile direction $Q_1$ and $Q_3$ as any solution of

$$\int_{\phi-Q_1}^{\phi} f(\theta)d\theta = 0.25 \text{ and } \int_{\phi}^{\phi+Q_3} f(\theta)d\theta = 0.25,$$

respectively. In most cases, the circular distribution is unknown. To date, no published literature is found on a nonparametric estimator of $Q_1$ and $Q_3$ for circular variables. However, it seems sensible to estimate $Q_1$ and $Q_3$ by classifying the sample observations into two groups based on their locations with respect to the sample median direction. Subsequently, $Q_1$ can be considered as the median of the first group and $Q_3$ as the median of the second.

Defining $Q_1$ and $Q_3$ in such a way may looks trivial because in some cases $Q_1$ could be larger than $Q_3$. Under this circumstance we can interchange the labels of $Q_1$ and $Q_3$ due to the compactness and periodicity of the circle. For simplicity and to avoid the confusion caused by the localization of $Q_1$ and $Q_3$, rotatable property of circular data by subtracting the estimated mean direction of the circular sample from each sample observation is used to make sure that the mean is in the 0 direction. This rotation might be helpful to identify $Q_1$ and $Q_3$ in a more consistent way. Hence, we can assume that $Q_1 \in (0, \pi]$ and $Q_3 \in (\pi, 2\pi)$. The robustness of mean direction, (see Wehrly and Shine(1981)) is a useful property which gives a fair assurance that the existence of any possible outlier will not have much effect on the estimated mean direction.

### 2.3 Circular interquartiles range $IQR$ and fences

Analogues to the linear case, interquartiles range $IQR$ is required to construct the circular boxplot. After rotation of sample observation, $IQR$ can be obtained by the following formula:

$$IQR = 2\pi - Q_3 + Q_1.$$

The upper and lower fences can be identified such as, lower fence $L_F = Q_1 + \nu IQR$ and upper fence $U_F = Q_3 - \nu IQR$, where $\nu$ is a resistance constant.

After obtaining the median,$Q_1, Q_3, L_F$ and $U_F$, it is possible to rerotate the sample by adding the estimated mean direction for rotated sample. Figure 1 illustrates the suggested shape of circular boxplot after the rotation. In the following section, numerical and simulation studies will be carried out in order to specify appropriate values of $\nu$.

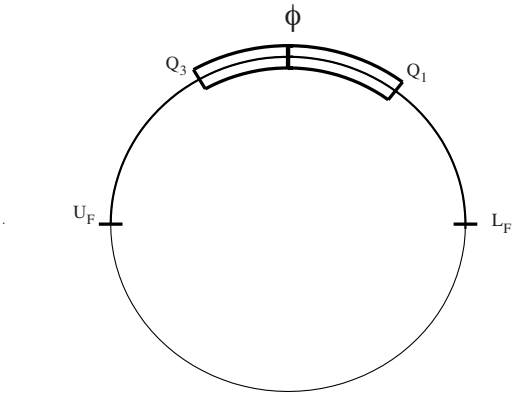


Figure 1: The proposed structure of circular boxplot.

## 3 Estimation of the resistance constant $\nu$

In linear case, $1.5IQR$ criterion is a popular method to identify outliers. The interest of investigating the appropriate values of constant $\nu$ was developed since the first construction of boxplot. Researchers like Hoaglin et al.(1986), Ingelfinger et al.(1983) and Sim et al.(2005) discussed the appropriate values of $\nu$ which can be used to identify the outlier in linear samples. It is not sensible to utilize similar constants $\nu$ of linear boxplot in the case of circular due to the bounded range of the circle. Hence, there is high possibility of overlapping problem between lower and upper fences for large resistance constant $\nu$ when the concentration parameter $\kappa$ is small.

Hoaglin et al.(1986) used different measures to investigate the behavior of boxplot. In this paper, we employ one of the measurements to estimate an appropriate values of the constant $\nu$.

### 3.1 Simulation and numerical studies

In order to investigate the behavior of circular variables with respect to five different summaries which are the median, $Q_1, Q_3, L_F$ and $U_F$, series of simulation studies are carried out. Samples were generated from von Mises distribution $VM(\mu, \kappa)$ with different sizes between 5 and 200. Different values of concentration parameter $\kappa$ are considered, $\kappa = 0.5, 1(1)10$. Further, various values of the resistance constant $\nu$= 1(0.2)3 and 3.5 are utilized in order to obtain $L_F$ and $U_F$. The $B(\nu, n)$ denotes the probability that the von Mises sample of size $n$ contains no observation outside the interval $(L_F, U_F)$.

A total of 3000 samples were generated from von Mises distribution $VM(\mu, \kappa)$ for each combination of $n$ and concentration parameter $\kappa$. The outcomes of simulation studies are $IQR, Q_1, Q_3, L_F, U_F$ and $B(\nu, n)$.

Overlapping problem between the upper and lower fences is expected to occur for some values of $\nu$ because of the bounded range of circular variables. Such problems caused a messy structure of boxplot and it may leads to miss identification of outliers.

### 3.2 Description of $B(\nu, n)$ measure and discussion

Let $B(\nu, n)$ denotes the probability of no observation outside the interval $(L_F, U_F)$ for von Mises sample of size $n$ and resistance constant $\nu$. Overlapping problem affects the behavior of $B(\nu, n)$. It is noticed that, for small concentration parameter $\kappa$, $B(\nu, n)$ is non monotone function of $\nu$, while $B(\nu, n)$ is an increasing function of $\nu$ for large concentration parameter, $(\kappa \geq 3)$. Further, for large concentration parameter $(\kappa > 3)$, $B(\nu, n)$ is not much affected by the increment of $\kappa$, as shown in Figure 2.

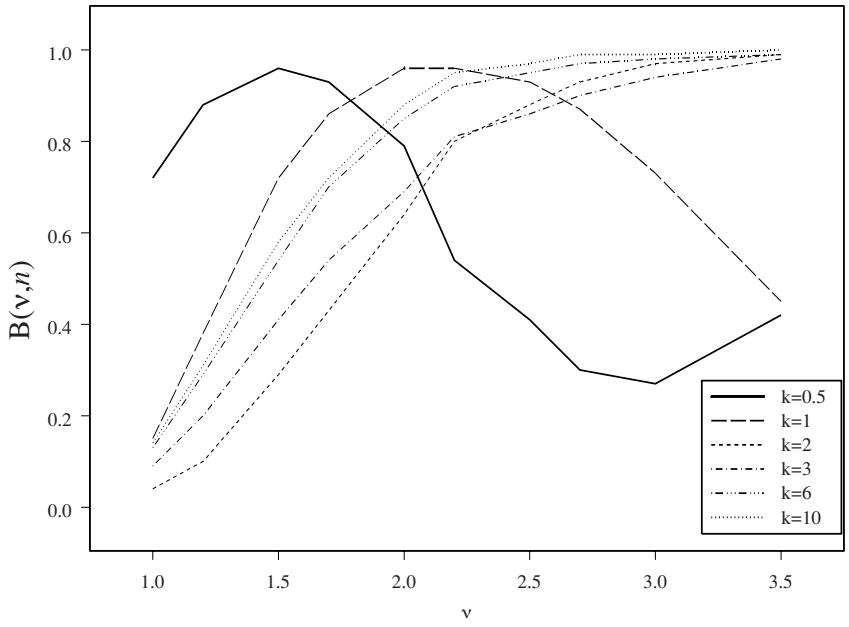


Figure 2: Behavior of $B(\nu, n)$ measure.

Simulation results of $B(\nu, n)$ are used to specify values of $\nu$ which can be implied on the structure of circular boxplot. It is more informative to interpret the results of simulation studies according to the mod of sample size $n$ with respect to 4, the sample size can be clustered into one of 4 groups according to whether $n$ has the form $4j, 4j+1, 4j+2$ or $4j+3$, where $j \in \mathbf{N}$.

Figure 3 shows the values of $\nu$ for sample size $n < 56$ at 0.1, 0.05 and 0.01 level of significance and large values of $\kappa$, $(\kappa = 7)$. It is shown that the values of $\nu$ are decreasing functions of the level of significance. At 0.05 level of significance, $\nu$ values seem to be stationary for $(5 < n < 130)$, with respect to the remainder after dividing sample size $n$ by 4. Results suggest that it is appropriate to use the value of resistance constant to be $(2.1 < v < 2.7)$. Simulation results show that the values of $\nu$ increase slightly for larger sample size. Similar behavior can be observed for $\alpha = 0.1$, but the values of resistance constant $\nu$ alternate between 1.5 and 2.2. The situation is different when $\alpha = 0.01$, where the cut points are fixed at $\nu = 3.5$ for $(5 < n < 25)$, and decreases to $\nu = 3$ for larger sample size $n$. For small concentration parameter $(\kappa < 3)$, the convenient values of resistance constant $\nu$ are to be less than 2.

## 4 Numerical example: Wind direction data

Abuzaid et al.(2008) have identified single outlier in linear circular regression model based on the circular residuals. They used several numerical and graphical methods to identify possible outliers in wind direction data set based on circular residuals. Two different techniques were used to measure wind directions along the Holereness coastline (the Humberside coast of North Sea, United Kingdom). a total of 129 observations were recorded in radian over 22.7 days. The data were fitted by using the simple regression model for circular variables which was proposed by Hussin et al.(2004). The fitted model for the data is given by:

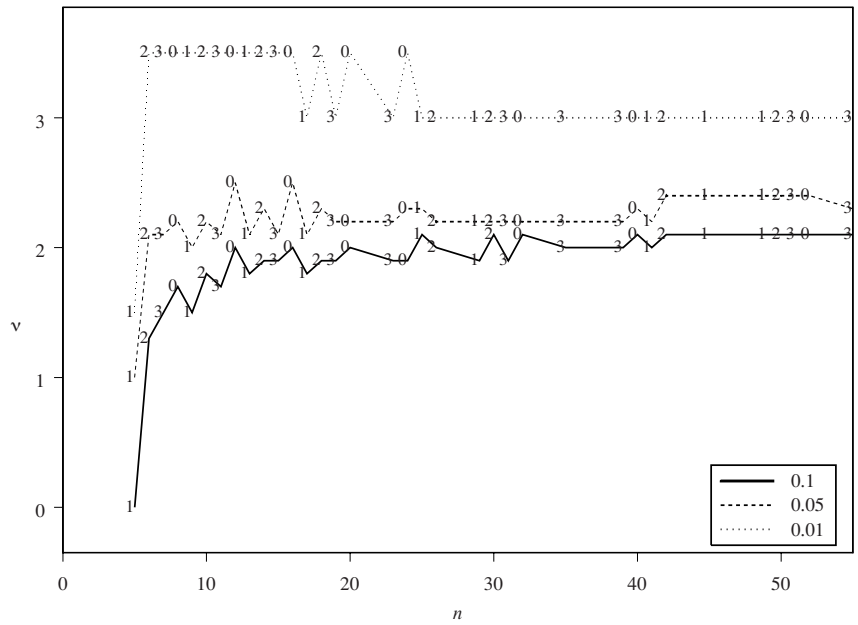


Figure 3: Percentile points for different sample size $n$, at $\kappa = 7$.

$$\hat{y}_i = 0.165 + 0.973x_i(mod 2\pi).$$

Then the residuals were estimated together with its mean direction ($\bar{\theta} = 0.017$), concentration parameter ($\hat{\kappa} = 7.34$),median direction ($\phi = 0.0072$), first quartile ($Q_1 = 0.202$), third quartile ($Q_3 = 6.125$) and ($IQR = 0.360$). By using numerical and graphical methods, Abuzaid et al.(2008) identified observations number 38 and 111 as outliers. Circular boxplot was used to identify possible outliers in circular regression via the circular residuals. Since $\hat{\kappa} = 7.34$, which is considered large enough, we can use the values of $\nu$ larger than 2.

Table 1 shows the observations detected as an outlier in wind direction data set by using different values of $\nu$, and the observation numbers 38 and 111 were identified as outliers for all values of $\nu$.

Table 1: Summary of the observations detected by using several values of $\nu$ for circular residuals of wind data.

| $\nu$ | $U_F$ | $L_F$ | **Count** | **Observation** |
|---|---|---|---|---|
| 1 | 0.561 | 5.765 | 14 | 15,18,38,43,48,68,70,95,98,99,100,109,111,123. |
| 1.2 | 0.633 | 5.693 | 12 | 18,38,43,48,68,70,95,98,99,100,111,123. |
| 1.5 | 0.741 | 5.586 | 6 | 38,43,70,99,100,111. |
| 1.7 | 0.813 | 5.514 | 4 | 38,43,70,111. |
| 2 | 0.921 | 5.406 | 3 | 38,43,111. |
| 2.2 | 1.029 | 5.298 | 2 | 38,111. |
| 2.5 | 1.101 | 5.226 | 2 | 38,111. |
| 2.7 | 1.173 | 5.154 | 2 | 38,111. |
| 3 | 1.281 | 5.046 | 2 | 38,111. |
| 3.5 | 1.461 | 4.866 | 2 | 38,111. |

For smaller values of $\nu$ many other observations are identified as outliers. Figure 4 shows the boxplot of circular residuals of wind direction data for $\nu$= 2.5.

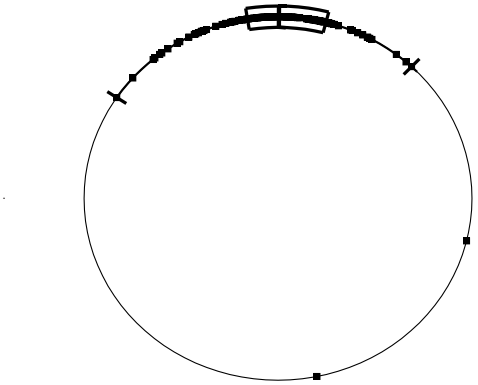

Figure 4: Circular boxplot of circular residuals of wind data, for $\nu$=2.5.

## 5 Conclusions

Boxplot is a popular tool for explanatory data analysis. It was developed gradually over the past 40 years and however there is no known structure of boxplot for circular variables. It is shown that by specifying the median direction, first and third quartiles solve half of the problem of constructing circular boxplot, while the determination of the upper and lower fences are more challengeable because of the bounded rang of the circle. further more the level of concentration parameter is highly affecting the structure of circular boxplot.

There are some interesting points being highlighted based on the simulation studies in Section 3, such as the robustness of mean direction in circular sample and the functional relationship between the $IQR$ and the large concentration parameter $\kappa$.

It is recommended to use different values of $\nu$ to identify possible outlier in circular variables. For samples with large concentration parameter $\kappa$, it is appropriate to use different values of $\nu$, where $2 < \nu < 2.7$, while for samples with small concentration parameter $\nu$, the values of resistant rules can be chosen between 1 and 2. These values are comparable to the linear case in which $\nu$ equal to 1.5.

Circular boxplot is also able to identify possible outlier in frogs' data set and the identification results are consistent with Collett(1980).Further, in the case of circular regression model, observation numbers 38 and 111 were identified as outliers by circular boxplot.

We have shown that the circular boxplot can be considered as a new graphical tool for explanatory analysis of circular data, and it is an alternative technique to identify outliers in circular samples and in circular regression through the circular residuals.

## References


Abuzaid, A. H., Hussin, A. G. and Mohamed, I. B. (2008). Identifying single outlier in linear circular regression model based on circular distance. *Journal of Applied Probability and Statistics*, 3(1): 107-117, Dixie W Publishing Corporation.

Barnett, V. and Lewis, T. (1978). *Outliers in Statistical Data*. New York and London,Wiley.

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

David, H. A. (1970). *Order statistics*. New York and London, Wiley.

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

Graedel, T. E.(1977). The wind boxplot: an improved wind rose. *Journal of applied meteorology*, 16: 448-450, American Meteorological Society.

Hoaglin, D. C., Iglewicz, B. and Tukey, J. W. (1986). Performance of some resistant rules for outlier labeling. *Journal of the American Statistical Association*, 81(396): 991-999, Alexandria, VA.

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 and Technology*, 8(1 & 2): 1-6, (INSS)and (ICMST).

Ingelfinger, J. A., Mosteller, F., Thibodeau, L. A., and Ware, J. H. (1983). *Biostatistics in Clinical Medicine*. New York, Macmillan.

Mardia, K. V.(1975). Statistics of directional data. *J. R. Statistic. Soc.*B, 37: 349-393, Blackwell.

Mardia, K. V. (1972). *Statistics of directional data*. Academic Press. London.

Mardia, K. V. and Jupp, P. E. (2000). *Directional data, 2nd edition*. J. Wiley. London.

Sim, C. H., Gan, F. F. and Chang, T. C. (2005). Outlier Labeling With Boxplot Procedures. *Journal of the American Statistical Association*, 100 (470): 642-652, Alexandria, VA.

Tukey, J. W. (1977). *Exploratory data analysis*. Addison-Wesley, Reading, MA.

Wehrly, T., Shine, E. P. (1981). Influence curves of estimates for directional data. *Biometrika*, 68: 334-335, Oxford University.