On Development of Spoke Plot for Circular Variables

Fakhrulrozi Hussain, Abdul G. Hussin*, and Yong Z. Zubairi
Centre for Foundation Studies in Science, University of Malaya, 50603 Kuala Lumpur, Malaysia.
*Author for correspondence; e-mail: ghapor@um.edu.my

ABSTRACT

Relationship and simple analysis between two any circular variables are the main focus of this paper. Correlation measure, linear relationship and estimation of concentration parameters are some examples of simple approach in analyzing two circular variables. These analyses could be extended to the graphical plots to give a better understanding. A plot called “Spoke plot” which have been developed in MATLAB environments have been proposed to achieve this objective. As an illustration, the application of “Spoke plot” as a simple analysis between two circular variables using graphical plots in investigating the correlation, linear relationship and estimating concentration parameters are given using the Malaysian wind direction data.

Keywords: concentration parameters, correlation coefficient, wind direction.

1. INTRODUCTION

Circular variables are common in some data such as clock time, compass bearing, moon circle, wind direction, turbine and others. The wind direction data is a type of the circular data. The data is measured on a scale that repeats itself or with angular directions. Unlike the common linear data, the features of circular data warrants appropriate techniques and special inferential tools in the analysis. The analysis of circular data has been studied for many years [1 - 4] as well as the application of the simple circular regression model was discussed [5]. Another area of circular statistics which is problems of outliers have been studied by Abu Zaid et al. [6]. Further, a number of symmetric probability models have been developed for directional data and these include the von Mises, wrapped Cauchy, Sen Gupta-Rattihalli and many others.

However, the analysis of directional data is limited as user friendly statistical software that deal with both exploratory data analysis as well as statistical inference are not many in the market. Although theoretical formulations in modeling circular variables have been developed, the evaluations are somewhat complicated due to the nature of the data. Therefore, any development on the analysis of circular data as well as to incorporate further analysis into the available software is indispensable for circular data. This paper focuses on graphical method in analyzing the relationship of two any circular variables via correlation and linear relationship.
2. CIRCULAR VARIABLES

Circular random variable is one which takes values on the circumference of a circle, i.e. they are angles in the range \((0, 2\pi]\) radians or \((0^\circ, 360^\circ]\). This random variable must be analyzed by techniques differing from those appropriate for the usual Euclidean type variables because the circumference is a bounded closed space, for which the concept of origin is arbitrary or undefined. A continuous linear variable is a random variable with realizations on the straight line or real line which may be analyzed straightforwardly by usual techniques. For example, if one wants to compare the difference between two circular data that have value \(10^\circ\) and \(350^\circ\), by using linear technique will give the answer of \(340^\circ\); but in reality, the difference between these two data is only \(20^\circ\) which suggests the need of special approach in dealing with such variables.

2.1 Correlation coefficient of circular variables

The correlation measure will be used in the development of the analysis of two circular variables. Correlation or also known as a measure of a correlation coefficient indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two variables from independence. When the data are linear there are several coefficients, measuring the degree of correlation, adapted to the nature of data.

Given \(n\) pairs of circular data \((\theta_1, \phi_1), \ldots, (\theta_n, \phi_n)\), where \(0 < (\theta, \phi) \leq 2\pi\), the circular correlation coefficient \([3]\) is defined by

\[
\rho_f = \frac{\sum_{1 \leq i, j \leq n} \sin(\theta_i - \theta_j) \sin(\phi_i - \phi_j)}{\sqrt{\sum_{1 \leq i, j \leq n} \sin^2(\theta_i - \theta_j) \sum_{1 \leq i, j \leq n} \sin^2(\phi_i - \phi_j)}},
\]

where \(-1 \leq \rho_f \leq 1\).

2.2 Linear relationship between two circular variables

The regression model when both variables are circular produces is a very interesting form. The model is given by \(\phi = \alpha + \beta \theta + \varepsilon \pmod{2\pi}\), where \(\varepsilon\) is a circular random error having a von Mises distribution with circular mean \(\mu\) and concentration parameter \(\kappa\), which can be written as \(\varepsilon \sim \text{VM}(\mu, \kappa)\).

This model have been discussed in detail for \(\beta = 1\) \([7]\) and also for \(\beta\) close to unity \([8]\). The estimates of \(\alpha, \beta\) and \(\kappa\) namely \(\hat{\alpha}, \hat{\beta}\) and \(\hat{\kappa}\) are given by

\[
\begin{align*}
\hat{\alpha} = & \begin{cases} 
\tan^{-1}\left(\frac{S}{C}\right), & S > 0, C > 0 \\
\tan^{-1}\left(\frac{S}{C}\right) + \pi, & C < 0 \\
\tan^{-1}\left(\frac{S}{C}\right) + 2\pi, & S < 0, C > 0
\end{cases}
\end{align*}
\]

where \(S = \sum \sin(\phi - \beta \theta)\) and \(C = \sum \cos(\phi - \beta \theta)\). Due to the nonlinear nature of the first partial derivative of the log likelihood function, then \(\beta\) is obtained by iterative procedure according to the formula

\[
\hat{\beta} \approx \hat{\beta}_0 + \frac{\sum \theta_i \sin(\phi_i - \hat{\alpha} - \hat{\beta} \theta_i)}{\sum \theta_i^2 \cos(\phi_i - \hat{\alpha} - \hat{\beta} \theta_i)}
\]

The estimate of \(\kappa\) is given by

\[
\hat{\kappa} = \frac{1}{n} \sum \cos(\phi_i - \hat{\alpha} - \hat{\beta} \theta_i)
\]

which can be approximated by

\[
A^{-1}(w) \approx \frac{9 - 8w + 3w^2}{8(1 - w)}
\]

3. THE DEVELOPMENT OF SPOKE PLOT

Two circular variables \((\theta, \phi)\) may be represented in a Spoke plot of two concentric non intersecting circles with any specific radius. To understand this idea, suppose we have a pair of observation \((45^\circ, 90^\circ)\) for variable \(\theta\) and \(\phi\) respectively. In the Spoke plot these two observations may be represented by a
line that connects these two points $\theta$ and $\phi$ from inner circle (45°) to outer circle (90°) respectively as shown in Figure 1. The choice of points for the inner and outer circle is arbitrary; that is $\theta$ may very well be chosen as the points in the inner and $\phi$ the points on outer circle respectively. Based on this idea, a Spoke plot is developed that allows us to look at the pattern of any two circular variables. Therefore, given a set of data with size $n$ where the coordinates $(\theta_i, \phi_i)$ for $i = 1, 2, \ldots, n$ are approximately equal (i.e. $\theta_i \approx \phi_i \mod 2\pi$) or linearly related, the set of $n$ lines formed by connecting points from the inner circle to outer circle results in a spoke-like image. This pattern suggests a linear association between the two circular variables. As an illustration, we generated values of $\theta$ and $\phi$ from the von Mises distribution with mean $\pi$ and concentration parameter 2 for both $\theta$ and $\phi$. A sample measure in degree of the simulated data is given in Table 1.

<table>
<thead>
<tr>
<th>$\theta$</th>
<th>3.56</th>
<th>104.69</th>
<th>72.28</th>
<th>104.51</th>
<th>321.22</th>
<th>93.11</th>
<th>348.57</th>
<th>63.62</th>
<th>152.67</th>
<th>13.88</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\phi$</td>
<td>110.15</td>
<td>85.40</td>
<td>86.24</td>
<td>51.55</td>
<td>223.60</td>
<td>19.96</td>
<td>100.48</td>
<td>16.98</td>
<td>44.51</td>
<td>0.90</td>
</tr>
</tbody>
</table>

The Spoke plot may be used to represent the pair of points graphically. A program called “Spoke(spoke_data)” is developed with output as shown in Figure 2. The Spoke plot in Figure 2 could be an alternative approach to look at the relationship between two circular variables $\theta$ and $\phi$ using diagrammatical representation. As for comparison to linear, a scatter plot that is normally used in linear analysis have been plotted as shown in Figure 3. The scatter plot, however, could be misleading due to the wrap around properties of the circular data and the Spoke plot seems to be the better alternative.
3.1 Spoke plot: Correlation coefficient

Another application of Spoke plot are for graphical representation for correlation coefficient and linear association. For the illustration purpose, a dataset called “spoke_data” is generated with size $n = 30$ based on a simple circular regression model, which is $\phi = \alpha + \beta \theta + \epsilon \pmod{2\pi)}$. Without loss of generality, $\alpha = 0$ and $\beta = 1$ have been chosen. Variable $\theta$ have been generated from $\mathcal{V}(\pi/4, 1.5)$ and $\epsilon$ from $\mathcal{V}(0, 30)$ respectively. The correlation coefficient $\rho_\parallel$ is used to measure of linear association between those two circular variables using a subprogram called “corr(spoke_data)”. The subprogram is designed in such a way that the first output gives a graphical view of the relationship of two circular variables. This is followed by a numerical measure that describes the strength of linearity. As an illustration the call function “spoke(spoke_data)” have been executed followed by call function “corr(spoke_data)”. From the output as shown in Figure 4, the Spoke plot strongly suggests the presence of linear association between two circular variables. This finding is further supported by a numerical measure of correlation value equals to 0.9555 as shown at the bottom of the plot.
3.2 Spoke plot: linear relationship using simple circular regression

Once a linear association is established from the graphical Spoke plot and supported by a numerical measure of $\rho_T$, the next step of the analysis is to estimate the parameters of linear relationship of $\varphi = \alpha + \beta \theta + \epsilon \ (\text{mod} \ 2\pi)$, namely $\alpha$ and $\beta$ respectively. Another subprogram called “alpha_regress (spoke_data)” is written to give the estimated value of $\alpha$ and $\beta$ using the maximum likelihood estimation. By using the similar dataset, we found that the estimation parameters are $\alpha = 6.2792$ radian and $\beta = 0.9981$ respectively as shown in Figure 5. These estimated values are very close to the true values used in the simulation.

Another applicability of the Spoke plot...
is to visually assess the concentration parameter \( \kappa \) of the data. A subprogram called “kappa_spoke(spoke_data)” have been written to estimate and display the concentration parameter for both \( \theta \) and \( \phi \) which are \( \kappa_{\theta} \) and \( \kappa_{\phi} \) respectively. In summary, the four commands given below produce a comprehensive output of spoke plot as shown in Figure 6.

\[
\begin{align*}
> & \text{spoke(spoke_data)} \\
> & \text{corr(spoke_data)} \\
> & \text{alpha_regress(spoke_data)} \\
> & \text{kappa_spoke(spoke_data)}
\end{align*}
\]

The program and subprograms to obtain Spoke plot for circular variables are available at URL: http://asasi.um.edu.my/download/Spoke_plot_program.doc.

![Figure 6. A sample of comprehensive Spoke plot.](image)

4. APPLICATION OF SPOKE PLOT

As an illustration, the Spoke plots are used to two different real dataset as described below:

4.1 wind direction data recorded at 850 hectopascals (hPa) and 1,000 hectopascals (hPa) at time 12.00A.M. from Bayan Lepas Airport, Malaysia, in July and August 2005, with the objective to compare two sets of wind direction data at two different pressures.

4.2 wind direction data from the Holderness Coastline, which is the Humberside coast of the North Sea, United Kingdom in October 1994 that were measured by HF radar and anchored wave buoy. The dataset have 49 measurements recorded over the period 22.7 days.

The results are shown in Figures 7 and 8 respectively. From the Spoke plot in Figure 7 of data set (i), it can be seen that a number of lines crossing the inner ring implies that there is no correlation between the variables. To support the finding, the calculated correlation value is 0.1316 which indicates a very weak correlation. The linear relationship value, however shows a strong of one to one linear relationship between the two circular variables. The concentration parameter estimate shows a small concentration for both \( \theta \) and \( \phi \).
Figure 7. Spoke Plot of wind direction data recorded at 850 hPa and 1,000 hPa, from Bayan Lepas Airport, Malaysia, in July and August 2005.

Figure 8. Spoke plot of wind direction measured by HF radar and anchored wave buoy, Holderness Coastline, Humberside coast of the North Sea, U.K., in October 1994.

From the Spoke plot in Figure 8 of data set (ii), it can be seen that none of the line crosses the inner ring which indicates a strong correlation between the variables. To support the finding, the calculated correlation value of 0.8680 is obtained. The linear association measure shows a strong one to one linear relationship between the two circular variables. Further, concentration parameters show a small concentration for both $\theta$ and $\phi$. 
5. DISCUSSION

In this paper, a pragmatic approach in the exploratory data analysis of circular variables is developed. The novelty of the methodology is that it combines both visual methodology and numerical approach in the analysis. With the name Spoke plot, the analysis focuses on identifying the relationship of two circular variables. Plot of Spoke graph, correlation measure, linear relationship and concentration parameter value provides a comprehensive exploratory data analysis. In summary, the method of analysis developed in this study has great potential being developed into comprehensive statistical software dedicated to circular variables. With the proper interface, option like graphical-user-interface (GUI), point-and-click window and multiple windows can make the software as sophisticated as those designed for linear data.

REFERENCES