The circular statistics toolbox for Matlab

Philipp Berens¹ and Marc J. Velasco²

April 2009

This Technical Report has been approved by:

Director at MPIK               Postdoc at MPIK
The circular statistics toolbox for Matlab

Philipp Berens$^1$ and Marc J. Velasco$^2$

April 2009

$^1$Computational Vision and Neuroscience Group, Max Planck Institute for Biological Cybernetics, Tübingen, Germany, email: berens@tuebingen.mpg.de

$^2$Center for Complex Systems and Brain Sciences, Florida Atlantic University, Boca Raton, USA, email: velasco@ccs.fau.edu

Abstract. Directional data is ubiquitous in science and medicine. Due to its circular nature such data cannot be analyzed with commonly used statistical techniques. Despite the rapid development of specialized methods for directional statistics over the last thirty years, there is not a lot of software available that makes such methods easy to use for users. Most importantly, one of the most commonly used programming languages in biosciences, Matlab, is currently not supporting directional statistics. To remedy this situation, we have implemented a toolbox in Matlab, which covers diverse aspects of the statistical analysis of directional data.

1 Introduction

Directional data is ubiquitous in various fields of science, as diverse as ocean research [4], agricultural science [1], psychology [13], medical research [10, 14], neuroscience [17] or zoology [3, 7, 15, 5], to name just a few. Consider as an example the study of Le et al., who analyzed the influence of the day of birth on the development of recurrent middle ear infection in children [14]. The scale “day of the year” is periodic in nature: It is of no concern whether a child is born on January 15th of 1989 or 1990 — for the purpose of the analysis pooled over years these two dates are the same. We can make this structure explicit by converting data of this type such as day of the year, time of the day or month of the year to angular directions as a common scale by

\[ x = \frac{2\pi y}{k}, \]

where \( y \) is the representation of the data in the original scale, \( x \) is its angular direction and \( k \) is the total number of steps on the scale that \( y \) is measured in. In the example, we have \( y \) representing day of the year and thus \( k = 365 \), not considering leap years for simplicity. In addition to data that needs to be converted to angular directions, measurements such as wind [4] or heading direction [7] measured using a compass come naturally in angular units. Instead of multiplying by \( 2\pi \) we can equivalently multiply by \( 360^\circ \) to obtain degrees instead of radians:

\[ x_{\text{degree}} = 360^\circ \cdot \frac{x_{\text{radians}}}{2\pi} \]  

(1)

Circular data measured only between 0 and 180 degrees may be converted to span the full directional scale by similar means. The analysis of directional data represents a particular challenge: There is no reason to designate any particular point on the circle as zero, just as it is somewhat arbitrary which day we consider the first in a new year. In addition, there is no inherent sense of high or low values. Therefore, ordinary statistics may generally not be used to draw conclusions from circular data. Consider for example calculating the mean of a set of three angles, 10°, 30° and 350°. The arithmetic mean of the angles is clearly 130°, pointing southeast, while all data samples point in northerly directions.

Circular or directional statistics is a subfield of statistics, which is devoted to the development of statistical methods suited for the analysis of directional data. While the field has been steadily developing since the early 1950es [9, 18], there are only few software packages available that offer a comprehensive repertoire of circular statistics methods to researchers. The circular package [16] for the R programming language as well as the circstats package [6] for Stata are both available as open-source and offer a variety of functions to users of these programming languages. In addition, the commercially available Oriana software [12], provides basic analysis functionality for Windows users. In biomedical research, however, the Matlab programming language is most widely used for data analysis. Yet at the same time, statistical packages or methods for the analysis of circular data are missing. To remedy this short-coming, we have developed the CircStat2009 toolbox for circular statistics for Matlab.
the use with Matlab, providing a comprehensive set of functions for the most common problems in descriptive and inferential statistics.

The aim of this report is not to provide a comprehensive review of circular statistics, but to describe the functionality of the presented toolbox. Thus, we will focus on the key elements of circular statistics necessary to understand and use the toolbox, while frequently referring to the existing literature for a more detailed discussion and examples.

2 Description of CircStat2009 functions

The CircStat2009 toolbox offers a variety of functions for performing descriptive and inferential statistics on directional datasets. All functions take data in radians as arguments. If no special reference is given, the described computations are more or less standard in directional statistics and can be found in any of the textbooks [11, 19, 8].

2.1 Descriptive statistics

circ\_mean\( \) computes the mean direction of a sample \( \alpha = (\alpha_1, \ldots, \alpha_N) \) consisting of \( N \) directional observations \( \alpha_i \). To this end, directions are first transformed to unit vectors in the two-dimensional plane by

\[
  r_i = \begin{pmatrix}
    \cos \alpha_i \\
    \sin \alpha_i
  \end{pmatrix}.
\]

After this transformation, the vectors \( r_i \) are vector averaged by

\[
  \bar{r} = \frac{1}{N} \sum_i r_i.
\]

The vector \( \bar{r} \) is called mean resultant vector. To yield the mean angular direction \( \bar{\alpha} \), \( \bar{r} \) is transformed using the four quadrant inverse tangent function. For ease of implementation in Matlab, the first transformation is implemented exploiting the identity

\[
  \cos \alpha + i \sin \alpha = \exp(i\alpha)
\]

and the second transformation uses the Matlab function angle.

If supplied with a second optional input argument \( w \) of the same length as \( \alpha \), circ\_mean\( \) assumes that the data has been binned with bin center \( i \) equal to \( \alpha_i \) and \( w_i \) equal to the number or fraction of samples falling into bin \( i \). If a second and third output argument are requested, circ\_mean\( \) computes the 95\% confidence intervals on the estimation of \( \bar{\alpha} \) using circ\_meanconf.

circ\_median\( \) computes the median direction \( \hat{\alpha} \) of a sample \( \alpha \). To compute the median, the diameter of the circle that devides the data into two equal sized groups is found. The median is the endpoint of the diameter closer to the center of mass of the data. If \( n \) is even, it lies halfway between the two closest datapoints. If \( n \) is odd, it falls on one of the data points. This function does not take binned data for obvious reasons.

circ\_r\( \) computes the length of the mean resultant vector \( \bar{r} \) as described above by

\[
  R = \|\bar{r}\|.
\]

Again, a second optional argument \( w \) allows for the treatment of binned data. However, the estimation of \( R \) is biased when binned data is used. This bias can be corrected for by supplying the bin spacing \( d \) as a third optional argument and computing a correction factor [19, equ. 26.16]

\[
  c = \frac{d}{2 \sin(d/2)},
\]

setting \( R_c = cR \).

The length of the mean resultant vector is a crucial quantity for the measurement of circular spread or hypothesis testing in directional statistics.
**circ_var** computes the circular variance defined as

\[ S = 1 - R. \]

Note that the circular variance \( s \) lies in the interval \([0, 1]\). If all samples point into the same direction, the resultant vector will have length close to 1 and the circular variance will correspondingly be small. If the samples are spread out evenly around the circle, the resultant vector will have length close to 0 and the circular variance will be close to maximal. On the other hand, a circular variance of 1 does not imply a uniform distribution around the circle. Optionally, circ_var can be used with binned data (see discussion for circ_r).

**circ_std** computes the angular deviation defined as

\[ s = \sqrt{2(1 - R)}. \]

Note that the angular deviation lies in the interval \([0, \sqrt{2}]\). Alternatively, the function computes the circular standard deviation

\[ s_0 = \sqrt{-2 \ln R} \]

ranging from 0 to \( \infty \). Generally, the first measure is preferred, as it is bounded, but the two measures deviate little [19]. Optionally, circ_std can be used with binned data (see discussion for circ_r).

**circ_axialmean** computes both the direction and the length of the of the mean resultant vector \( \vec{r} \) for axial (or multi-modal) data, using the form

\[ \vec{r}_m = \frac{1}{N} \sum_k \exp (im\alpha_k). \]

In this case, the mean direction \( \bar{\alpha}_m = \frac{1}{m} \text{angle}(\vec{r}_m) \) and the length \( r_m = ||\vec{r}_m||. \)

This method is the same as computing the resultant vector, except that angles are multiplied by the number of modes \( m \) (for axial data \( m = 2 \)). The effect of the multiplication is to rotate vector components that would normally cancel out so that they instead overlap. For example, consider a sample with two modes, one at 0, and the other near \( \pi \). Normally, the mean resultant would have zero length and arbitrary direction. After multiplying by \( m = 2 \), the components near 0 are still there, and the components at \( \pi \) have been rotated over to \( 2\pi \) (which modulo \( 2\pi \) is 0). The components overlap and the statistics of the resultant vector more adequately captures the structure of the data. See also: [2, equations 1.6.1, 2.3.5, 17.2.7].

**circ_moment** computes the p-th trigonometric moment of a sample. The mean direction and the mean resultant length (described above) are the angular and amplitude component of the first trigonometric moment. Higher trigonometric moments are useful in calculations of skewness and kurtosis. In general, the p-th moment can be defined as [8, equation 2.20]

\[ m_p = \bar{C}_p + i\bar{S}_p = R_p e^{i\bar{\alpha}_p} \]

where

\[ \bar{C}_p = \frac{1}{n} \sum_{i=1}^{n} \cos p\alpha_i, \quad \bar{S}_p = \frac{1}{n} \sum_{i=1}^{n} \sin p\alpha_i. \]

**circ_corrcc** computes a correlation coefficient \( \rho_{cc} \) between two directional random variables \( \alpha \) and \( \beta \) [11, p. 176] by

\[ \rho_{cc} = \frac{\sum_i \sin(\alpha_i - \bar{\alpha}) \sin(\beta_i - \bar{\beta})}{\sqrt{\sum_i \sin^2(\alpha_i - \bar{\alpha}) \sin^2(\beta_i - \bar{\beta})}}, \]

where \( \bar{\alpha} \) denotes the mean direction computed by circ_mean. A p-value for \( \rho_{cc} \) is computed by considering the test statistic

\[ t = \sqrt{f} \cdot \rho_{cc}, \]

which is distributed as standard normal under the null hypothesis of zero correlation. The term \( f \) is given by

\[ f = N \sum_i \sin^2(\alpha_i - \bar{\alpha}) \sum_i \sin(\beta_i - \bar{\beta}) \sum_i \sin^2(\alpha_i - \bar{\alpha}) \sin^2(\beta_i - \bar{\beta}). \]
**circ_correl** computes a correlation coefficient $\rho_{cl}$ between a directional random variable $\alpha$ and a linear variable $x$, by correlating $x$ with $\cos \alpha$ and $\sin \alpha$ individually. To this end, we define the correlation coefficients $r_{sx} = c(\sin \alpha, x), r_{cx} = c(\cos \alpha, x)$ and $r_{cs} = c(\sin \alpha, \cos \alpha)$, where $c(x, y)$ is the Pearson correlation coefficient. Then the circular-linear correlation is given by [19, equ. 27.47]

$$
\rho_{cl} = \sqrt{r_{sx}^2 + r_{cx}^2 - 2r_{cx}r_{sx}r_{cs}} / (1 - r_{cs}^2).
$$

A p-value for $\rho_{cl}$ is computed by considering the test statistic $N\rho^2$, which $\chi^2$-distributed with 2 degrees of freedom.

**circ_confmean** compute the $(1 - \delta)$%-confidence intervals for the population mean [19, equations 26.23-26.26]. For $R \leq 0.9$ and $R > \sqrt{\chi^2_{\delta,1}/2N}$, we compute

$$
d = \arccos \left( \sqrt{\frac{2N(2R_n^2 - N\chi^2_{\delta,1})}{4N - \chi^2_{\delta,1}}} \right),
$$

where $R_n = R \cdot N$. For $R > 0.9$, we compute

$$
d = \arccos \left( \sqrt{\frac{N - (N^2 - R_n^2) \exp(\chi^2_{\delta,1}/N)}{R_n}} \right).
$$

In both cases, the lower confidence limit of the mean is found by $L_1 = \bar{\alpha} - d$ and the upper confidence limit by $L_2 = \bar{\alpha} + d$.

### 2.2 Inferential statistics

**circ_rtest** performs the Rayleigh test for circular uniformity. It asks how large the resultant vector length $R$ must be to indicate a non-uniform distribution. Under the null hypothesis $H_0$ the population is uniformly distributed around the circle. Under the alternative hypothesis $H_A$ the population is not distributed uniformly around the circle. Importantly, the Rayleigh test assumes that the sample is generated from a von Mises distribution (see below). The approximate p-value under $H_0$ is computed as [19, equation 27.4]

$$
P = \exp \left[ \sqrt{(1 + 4N + 4(N^2 - R_n^2) - (1 + 2N)} \right],
$$

where again $R_n = R \cdot N$.

**circ_otest** performs the “omnibus test” or Hodges-Ajne test for circular uniformity as an alternative to the Rayleigh test [19, equation 27.7-27.9]. It works well for unimodal, biomodal and multimodal distributions and detects therefore general deviations from uniformity. The null and alternative hypothesis are the same as for the Rayleigh test. If the conditions of the Rayleigh test are met, it is more powerful than the Hodges-Ajne test.

To conduct the test, the smallest number $m$ that occur within a range of 180 degree is computed. Under the null hypothesis, the probability of observing an $m$ this small or smaller is

$$
P = \frac{1}{2N-1} (N - 2m) \left( \begin{array}{c} N \\ m \end{array} \right),
$$

which can for $N > 50$ be approximated by

$$
P \approx \sqrt{\frac{2\pi}{A}} \exp \left( -\pi^2/(8A^2) \right),
$$

where $A = \frac{\pi \sqrt{N}}{2(N - 2m)}$.  

4
**circ_raoet** performs Rao’s spacing test for circular uniformity as an additional alternative to the Rayleigh test [2]. The null and alternative hypothesis are the same as for the Rayleigh test. It is more powerful than alternatives when the data are neither unimodal nor axially bimodal. It is based on the idea that in an ordered sample \( \alpha = (\alpha_1, \ldots, \alpha_N) \) with \( \alpha_{i+1} > \alpha_i \) sampled from a uniform distribution the differences between successive samples should be approximately \( \pm 360^\circ / N \). Its test statistic is defined as

\[
U = \frac{1}{2} \sum_{i=1}^{N} |d_i - \lambda|,
\]

with \( d_i = \alpha_{i+1} - \alpha_i \), \( d_N = 360^\circ - (\alpha_N - \alpha_1) \) and \( \lambda = \frac{360^\circ}{N} \) [2, equations 4.6.1-4]. The distribution of \( U \) is fairly complex, and we use tabled values instead of the full distribution.

**circ_vtest** performs the V test for circular uniformity which is similar to the Rayleigh test but where under the alternative hypothesis \( H_A \) the data is not uniformly distributed around the circle, but has a mean direction \( \alpha_A \). The test statistic is computed as [19, equation 27.5]

\[
V = R_n \cos(\bar{\alpha} - \alpha_A),
\]

where \( R_n \) as above. Approximate critical values for the quantity

\[
V \sqrt{\frac{2}{N}}
\]

can be obtained from the one tailed normal deviate \( Z_{\alpha/2} \). Note that not rejecting the null hypothesis in this case leaves it open whether the cause for that failure was insufficient evidence for non-uniformity or a different mean direction than \( \alpha_A \).

**circ_mtest** performs a one sample test asking whether the population mean is angle equal to a specified value. The null hypothesis \( H_0 \) is that the mean angle is equal to some value \( \bar{\alpha}_0 \) and the alternative hypothesis \( H_A \) is that this is not the case, i.e. \( \bar{\alpha} \neq \bar{\alpha}_0 \). The test at significance level \( \delta \) is performed by checking whether \( \bar{\alpha}_0 \in [L_1, L_2] \), where \( L_1 \) is the lower \( 1-\delta \) confidence limit on the population mean and \( L_2 \) the upper confidence limit.

**circ_wtest** performs the Watson-Williams two- or multi-sample test of the null hypothesis \( H_0 \) that any of \( s \) samples share a common mean direction, i.e. \( \bar{\alpha}_1 = \cdots = \bar{\alpha}_s \). The test statistic is calculated via [19, equation 27.14]

\[
F = K \frac{(N-s) \left( \sum_{j=1}^{s} R_j - R \right)}{(s-1) \left( N - \sum_{j=1}^{s} R_j \right)},
\]

where \( R \) is the mean resultant vector length across all data and \( R_j \) the mean resultant vector length computes on the \( j \)th samples alone. The correction factor \( K \) is computed from

\[
K = 1 + \frac{3}{8\kappa},
\]

where \( \kappa \) is the maximum likelihood estimate of the concentration parameter of a von Mises distribution with resultant vector length \( r_\nu \), which is the mean resultant vector length of the \( s \) resultant vectors \( r_j \). The obtained value of the test statistic is the compared to a critical value at the \( \delta \) level obtained from \( F_{\delta(1,1, N-s-2)} \).

The Watson-Williams test assumes underlying von Mises distributions with equal concentration parameter, but has proven to be fairly robust against deviations from these assumptions. The sample size for applying the test should be at least 5 for each individual sample. If binned data is used, bin widths should be no larger than 10 degree. Note that rejecting the null hypothesis only provides evidence that not all of the \( s \) samples come from a population with equal mean direction, not if all samples have pairwise differing mean directions or evidence of which of the samples differs.
**2.3 the von Mises distribution**

**circ_vmpdf**  The von Mises distribution is a probability density defined as

\[ f(\phi) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\phi - \theta)}. \]

The two parameters \( \kappa \) and \( \theta \) refer to the concentration and to the mean direction, respectively. \( I_0 \) is a Bessel function. The distribution is symmetric and unimodal. As \( \kappa \) approaches 0, the distribution approaches a uniform distribution on the circle. On the circle, the von Mises distribution is roughly analogous to the normal distribution. (See also: [2, equation 15.3.1])

**circ_ranvm**  Generates random angles drawn from a von Mises distribution. Parameters \( \theta \) and \( \kappa \) define the mean direction and concentration of the von Mises distribution to be sampled. If \( \kappa \) is 0, then samples are effectively drawn from a uniform distribution.

**circ_vmpar**  Given sample angles, this will find the maximum likelihood estimates of \( \hat{\kappa} \) and \( \hat{\theta} \) for the underlying population distribution. For von Mises distributions, the maximum likelihood estimate of the population direction equals the sample mean direction. Likewise, the maximum likelihood estimate of the concentration parameter equals the concentration parameter calculated from the sample mean vector length \( r \).

**circ_r2kappa / circ_kappa2r**  Convert between resultant length \( r \) and the corresponding best fit concentration parameter for a von Mises distribution \( \kappa \). The relation is given by

\[ A(\hat{\kappa}) = I_1(\hat{\kappa})/I_0(\hat{\kappa}) = r \]

where \( I_1 \) and \( I_0 \) are Bessel functions.

**2.4 Miscellaneous**

**circ_plot**  is a visualization tool for circular data. Supports three plotting styles: pretty, hist and regular. For a detailed documentation, see documentation in the file.

**circ_dist**  computes a vector of length \( N \) with the distances \( d_i \) between the paired samples \( \alpha = (\alpha_1, \ldots, \alpha_N) \) and \( \beta = (\beta_1, \ldots, \beta_N) \), where \( d_i = d(\alpha_i, \beta_i) \).

**circ_dist2**  computes a \( N \times M \) matrix with all pairwise distances \( d_{ij} \) of the two samples \( \alpha = (\alpha_1, \ldots, \alpha_N) \) and \( \beta = (\beta_1, \ldots, \beta_M) \), where \( d_{ij} = d(\alpha_i, \beta_j) \).

**ang2rad / rad2ang**  transform data from angular representation to radians and radians to angular representation, respectively, by using equation 1.

**circ_clust**  Performs a simple k-means clustering on the circle. First, cluster assignments are calculated based on the minimal distance from a cluster center. Next the two clusters with minimal distance between their centers are merged. This process is continued until the desired number of clusters is reached. Optionally displays its progress.

**3 Summary**

In this report, we described the CircStat2009 toolbox for performing statistical analysis of circular and directional data in Matlab. The functions cover a wide range of applications from descriptive statistics and inferential statistics. We supply the reader with parametric and nonparametric methods for testing a variety of hypothesis about circular data including ANOVA-like multisample testing. The goal of this toolbox is to make circular statistics methods available to a wider range of researchers, especially in applied fields of biomedical research.

While the most common applications and problems in directional statistics have been thoroughly covered, two extensions seem desirable: better support for ANOVA-like testing of multisample hypothesis in particular with regard to the common two-factor design and density estimation techniques on the for circular data. We will include these in future releases. The toolbox described in this document can be downloaded from http://www.mathworks.com/matlabcentral/fileexchange/10676.
Acknowledgments. The authors are grateful to M. Bethge, A. Tolias and A. Ecker for discussion. PB has been supported by the Max Planck Society, the German Ministry of Education, Science, Research and Technology through the Bernstein award to Matthias Bethge (BMBF; FKZ:01GQ0601) and a scholarship by the German National Academic Foundation.

References