Technical Report No. 184

The circular statistics toolbox for Matlab

Philipp Berens[1] and Marc J. Velasco[2]

April 2009

This Technical Report has been approved by:

Director at MPIK   Postdoc at MPIK

Max–Planck–Institut für biologische Kybernetik

Max Planck Institute for Biological Cybernetics

The circular statistics toolbox for Matlab

Philipp Berens[1] and Marc J. Velasco[2]

[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

The circular statistics toolbox for Matlab

Philipp Berens[1] and Marc J. Velasco[2]

Abstract. Directional data is ubiquitious 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 ubiquitious 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 y_i on the development of recurrent middle ear infection in children [14][1]. 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 leapyears 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 2π we can equvivalently multiply by $360°$ to obtain degrees instead of radians:

$$x_{degree} = 360° \cdot \frac{x_{radians}}{2\pi} \tag{1}$$

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 northernly 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 commerically available Oriana software [12], provides basic analysis functionality for Windows users. In biomedical reserach, 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

---

[1]For the curious reader it is mentioned that Le et al. indeed find an effect of day of birth on the development of middle ear infection: Children born during late winter or early spring are less likely to develop recurrent middle ear infection.

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 neccessary 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 α_i . To this end, directions are first transformed to unit vectors in the two-dimensional plane by

$$r_i = \left(\begin{array}{c} \cos \alpha_i \\ \sin \alpha_i \end{array} \right).$$

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 α, circ_mean assumes that the data has been binned with bin center i equal to α_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 α. 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 half-way 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.

2

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 ∞. 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 $\bar{r}$ for axial (or multi-modal) data, using the form

$$\bar{r}_m = \frac{1}{N} \sum_k \exp\left(im\alpha_k\right).$$

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

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 π. 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 π have been rotated over to 2π (which modulo 2π 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]

$$m_p = \bar{C}_p + i\bar{S}_p = R_p e^{i\bar{r}_p}$$

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 ρ_{cc} between two directional random variables α and β [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 ρ_{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

$$f = N \frac{\sum_i \sin^2(\alpha_i - \bar{\alpha}) \sum_i \sin(\beta_i - \bar{\beta})}{\sum_i \sin^2(\alpha_i - \bar{\alpha}) \sin^2(\beta_i - \bar{\beta})}.$$

3

circ_corrcl   computes a correlation coefficient ρ_{cl} between a directional random variable α 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]

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

A p-value for ρ_{cl} is computed by considering the test statistic $N\rho^2$, which χ^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 $R \le 0.9$ and $R > \sqrt{\chi_{\delta,1}^2/2N}$, we compute

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

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

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

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

$$P = \frac{1}{2^{N-1}}(N - 2m)\binom{N}{m},$$

which can for $N > 50$ be approximated by

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

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

circ_raotest 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 $\frac{360°}{N}$. Its test statistic is defined as

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

with $d_i = \alpha_{i+1} - \alpha_i$, $d_N = 360° - (\alpha_N - \alpha_1)$ and $\lambda = \frac{360°}{N}$ [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 $\bar{\alpha}_A$. The test statistic is computed as [19, equation 27.5]

$$V = R_n \cos(\bar{\alpha} - \bar{\alpha}_A),$$

where R_n as above. Approximate critical values for the quanitity

$$V \sqrt{\frac{2}{N}}$$

can be obtained from the one tailed normal deviate $Z_{\alpha(1)}$. 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 $\bar{\alpha}_A$.

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 δ 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_medtest performs a nonparametric test of the hypothesis H_0 that the median direction $\hat{\alpha} = \hat{\alpha}_0$ [19, section 27.3]. To this end, the number n of samples falling on either side of a diameter through $\hat{\alpha}_0$ are counted. The p-value for observing this sample under the null hypothesis is found by computing the probability of observing n or more out of N samples falling on one side of the diameter under a binomial distribution $B(N, p)$ with $p = 0.5$.

circ_wwtest 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 jth samples alone. The correction factor K is computed from

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

where κ is the maximum likelihood estimate of the concentration parameter of a von Mises distribution with resultant vector length r_w, 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 δ level obtained from $F_{\delta(1),1,N-2}$.

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.

circ_symtest   performs a simple test of symmetry of the sample around the median. Under the null hypothesis H_0 the underlying distribution is symmetrical around the median, while under the alternative hypothesis H_A it is not symmetrical around the median. To test this hypothesis, the circular distance d_i of each point from the median is computed. The d_i are then subjected to a Wilcoxon signed-rank test.

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 κ and θ refer to the concentration and to the mean direction, respectively. I_0 is a Bessel function. The distribution is symmetric and unimodal. As κ 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_randvm   Generates random angles drawn from a von Mises distribution. Parameters θ and κ define the mean direction and concentration of the von Mises distribution to be sampled. If κ 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 $\hat{\kappa}$. 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.

References

[1] Asa L. Aradottir, Alexander Robertson, and Edward Moore, *Circular statistical analysis of birch colonization and the directional growth response of birch and black cottonwood in south iceland*, Agricultural and Forest Meteorology **84** (1997), no. 1-2, 179–186.

[2] E. Batschelet, *Circular statistics in biology.*, Academic New York (1981).

[3] P. Berthold, A. J. Helbig, G. Mohr, and U. Querner, *Rapid microevolution of migratory behaviour in a wild bird species*, Nature **360** (1992), no. 6405, 668–670.

[4] J. A. Bowers, I. D. Morton, and G. I. Mould, *Directional statistics of the wind and waves*, Applied Ocean Research **22** (2000), no. 1, 13–30.

[5] William W. Cochran, Henrik Mouritsen, and Martin Wikelski, *Migrating songbirds recalibrate their magnetic compass daily from twilight cues*, Science **304** (2004), no. 5669, 405–408.

[6] Nicholas J. Cox, *Circular statistics in stata.*

[7] P. Domenici and R. W. Blake, *Escape trajectories in angelfish (Pterophyllum eimekei)*, J Exp Biol **177** (1993), no. 1, 253–272.

[8] N. I. Fisher, *Statistical analysis of circular data*, revised. ed., Cambridge Univ Pr, October 1995.

[9] Ronald Fisher, *Dispersion on a sphere*, Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences **217** (1953), no. 1130, 295–305.

[10] Fei Gao, Kee-Seng Chia, Ingela Krantz, Per Nordin, and David Machin, *On the application of the von mises distribution and angular regression methods to investigate the seasonality of disease onset*, Statistics in Medicine **25** (2006), no. 9, 1593–1618.

[11] S. Rao Jammalamadaka, Arjun Sengupta, and A. Sengupta, *Topics in circular statistics*, Har/Dsk ed., World Scientific Pub Co (, 2001.

[12] Computing Service Kovach, *Oriana*.

[13] Thomas Kubiak and Cornelia Jonas, *Applying circular statistics to the analysis of monitoring data*, European Journal of Psychological Assessment **23** (2007), no. 4, 227–237.

[14] Chap T. Le, Ping Liu, Bruce R. Lindgren, Kathleen A. Daly, and G. Scott Giebink, *Some statistical methods for investigating the date of birth as a disease indicator*, Statistics in Medicine **22** (2003), no. 13, 2127–2135.

[15] K Lohmann, N Pentcheff, G Nevitt, G Stetten, R Zimmer-Faust, H Jarrard, and L Boles, *Magnetic orientation of spiny lobsters in the ocean: experiments with undersea coil systems*, J Exp Biol **198** (1995), no. 10, 2041–2048.

[16] Ulric Lund and Claudio Agostinelli, *The circular package*.

[17] Pedro E. Maldonado, Imke Godecke, Charles M. Gray, and Tobias Bonhoeffer, *Orientation selectivity in pinwheel centers in cat striate cortex*, Science **276** (1997), no. 5318, 1551–1555.

[18] G. S. Watson and E. J. Williams, *On the construction of significance tests on the circle and the sphere*, Biometrika **43** (1956), 344–352.

[19] J. H. Zar, *Biostatistical analysis*, 4th ed., Prentice Hill, 1999.