ELSEVIER

Computers & Geosciences 32 (2006) 166–175

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

MATLAB functions to analyze directional (azimuthal) data—I: Single-sample inference ☆

Thomas A. Jones*

Department of Earth Sciences, Rice University, Houston, TX 77251-1892 USA

Received 11 September 2004; received in revised form 12 June 2005; accepted 13 June 2005

Abstract

Data that represent azimuthal directions cannot be analyzed statistically in the same manner as ordinary variables that measure, for example, size or quality. Many statistical techniques have been developed for dealing with directional data, but available software for the unique calculations is scarce. This paper describes a MATLAB script, located on the IAMG server, that calculates descriptive statistics and performs inference on azimuthal parameters. The calculations include tests for specific distributions, and tests on the preferred direction and concentration of vectors about this direction. The inference methods use large-sample approximations, plus resampling methods (bootstrap) for smaller sample sizes.

Keywords: Azimuths; Inference: Vector mean; Concentration; Von Mises; Uniform

1. Introduction

Data that describe directions are important in the earth sciences. The azimuthal direction of a fluvial current that forms cross-beds is an obvious example. However, observations from such directional variables cannot be analyzed with standard statistical methods. I use the geological definition of azimuths which are cyclical in nature, with 0° and 360° representing the same direction, due North. Azimuths increase clockwise, with 90° representing East. The cyclicity implies that if we observe directions of 2° and 358°, a simple mean would indicate a direction of 180°, or due South, whereas we know that the correct average or dominant direction is actually North. Special methods of analysis clearly are needed for such variables.

During the 1960s, statisticians and earth scientists started looking extensively at ways to statistically analyze azimuthal or other cyclical data. Statisticians have taken large steps in how to work with such variables in the past 30 years. Excellent references on cyclical variables, history of their analysis, and modern statistical methods include Mardia (1972), Fisher (1993), and Mardia and Jupp (2000). Applying some of these new techniques is difficult, requiring special code. Available software is scarce, but includes MATLAB scripts (Middleton, 2000), SAS macros (Blaesild and Granfeldt, 2003), code for displays (Wells, 1999), and the DDSTAP package developed by Ashis SenGupta (Indian Statistical Institute, Calcutta). I am not aware of general, inclusive computer programs that are designed for analyzing directional data.

MATLAB® (a registered trademark of The Mathworks, Inc.) is a general program for doing mathematical

☆Code available from IAMG server at http://www.iamg.org/CGEditor/index.htm

*Corresponding author at: 5211 Braeburn Drive, Bellaire, TX 77401-4814 USA.

E-mail address: tajones@houston.rr.com.

0098-3004/$ - see front matter © 2005 Elsevier Ltd. All rights reserved.
doi:10.1016/j.cageo.2005.06.009

computation, primarily designed to work with vectors and matrices. It is widely used in the engineering and physical sciences because of its rich set of capabilities. Many good references are available, including Hanselman and Littlefield (2001) and Hunt et al. (2001). Although it has some statistical capability, it is not a statistical-analysis tool. However, functions in MATLAB can be combined to execute complex statistical analyses.

MATLAB contains a strong capability for users to develop scripts and functions that make use of the basic MATLAB tools. The ability also exists to create graphic user interfaces (GUIs) for operation of the scripts. This paper describes a script that is designed for analysis of a sample of directional (azimuthal) data. This script, named **Vector_Stats**, calculates descriptive statistics, generates plots, and performs single-sample inference on distributions and parameters. A second paper describes another script (Jones, in press) that is designed for correlation analysis.

Vector_Stats, is designed strictly for two-dimensional directional data (that is, azimuths on maps), such as directions of cross-beds. Further, it is limited to vectors (e.g., direction of dip) that indicate a single direction. It is not appropriate for axial data (two-directional lines such as strike) that cannot distinguish between two directions 180° apart. See Section 4.2 for recommendations on dealing with axial data. This script is not for three-component directions, such as orientations of poles to bedding.

Although directions are the most common form of cyclical data in the earth sciences, other variables are also cyclical in nature and must be analyzed similarly to azimuths. For instance, the time of day that tornados occur is cyclical over 24 h, with 00:01 and 23:59 representing essentially the same time. Again, standard statistical techniques are not adequate for this type of data. Converting such variables to pseudo-angles (azimuths) allows analysis.

This paper describes methods that have been coded into the script. These are taken from the literature, primarily Fisher (1993) and Mardia and Jupp (2000). These books extensively reference prior work. Accordingly, readers are referred to the books for added information and original references. This paper's contribution is in selection and coding of the statistical methods for MATLAB.

2. Overview of orientation data

Assume we have measured a set of N cross-bed orientations, θ_i, each representing the direction observed at a location. The observations are in the form of azimuths, with 0° representing North, 90° East, and so on. Hence, the values of the θ_i are in the range (0°, 360°), or equivalently in the range of $(0, 2\pi)$ radians. We may think of each observation as a unit vector that points in the direction of interest.

Table 1 shows a listing of a typical data array used by the script. Each row in the array represents an observation. The columns represent different variables, some of which may be azimuthal and others linear. Here we are interested in column 1, which represents cross-bed orientations from the Bulgoo Formation, Australia. These data are provided on the IAMG server in file **TestData.txt**, and are discussed in Section 4.1.

Useful plots for showing the distribution of the observed data may be generated. Fig. 1A shows a rose diagram of the data in Table 1. Note that this plot is in terms of azimuths (i.e., North is 0 and angles increase in a clockwise direction). The MATLAB function **rose** generates a similar plot that is in terms of angles that follow the standard mathematical definition (that is, angles increase in a counter-clockwise direction). Wells (1999) points out the sensitivity of rose diagrams, similarly to histograms, to the origin and widths of the classes. He provides alternative methods, as well as QuickBasic code, for generating plots.

Table 1
Listing of typical data array used by **Vector_Stats**

294	1
177	2
257	3
301	4
257	5
267	6
329	7
177	8
241	9
315	10
229	11
239	12
277	13
250	14
287	15
281	16
166	17
229	18
254	19
232	20
290	21
245	22
245	23
214	24
272	25
224	26
215	27
242	28
186	29
224	30

Column 1 contains azimuths and column 2 is an identifier (Fisher, 1993, appendix B.6).

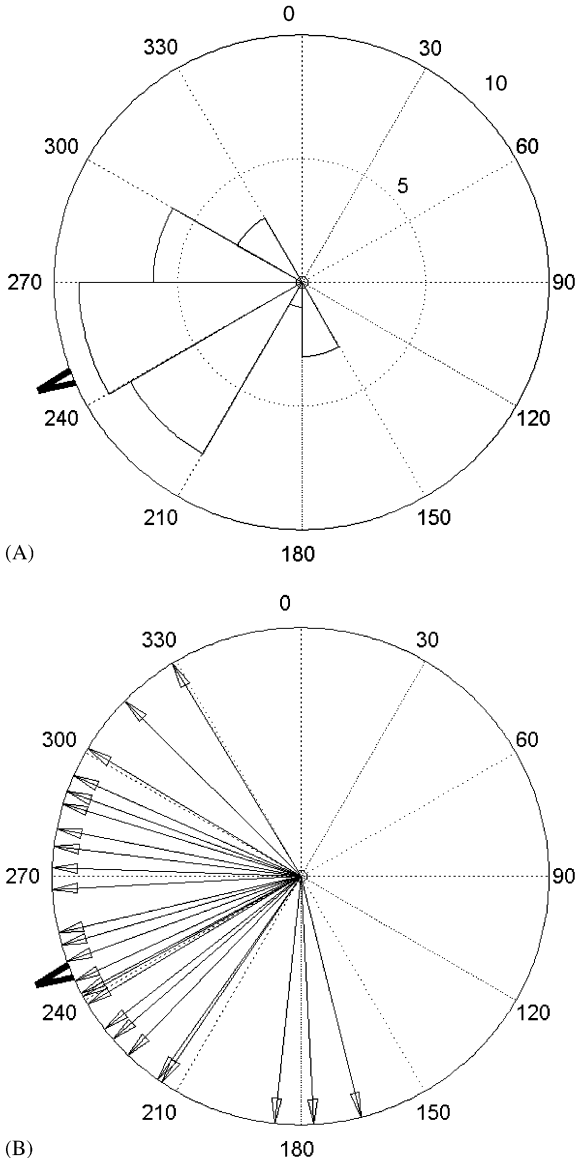

Fig. 1. Plots showing distribution of azimuths (Table 1) from the Bulgoo Formation, Australia (Fisher, 1993, appendix B.6); large arrowhead on outside of each circle shows direction of vector mean. (A) Rose diagram and (B) Compass plot.

Fig. 1B shows a compass plot that indicates the individual vectors, again plotted as azimuths. This is similar to the MATLAB function **compass**, which plots according to the mathematical orientation. Fisher (1993, chapter 2) shows several other ways to display observed azimuthal data.

2.1. Estimation of preferred direction

With ordinary linear data, we commonly calculate the sample mean to indicate the central portion of the observed data. Similarly with vector data, we wish to know the preferred orientation $\bar{\theta}$ of the sample (i.e., a measure that is analogous to the center of the observations but that takes cyclicity into account). An intuitive estimate of the vector mean (direction of preferred orientation) is given by the vector sum. That is, we add together the N unit vectors to form a combined vector, as in Fig. 2.

We can find the X and Y components of each vector by trigonometry, and combine the components to obtain the direction of the resultant vector. Define

$$C = \sum_{i=1}^{N} \cos\,\theta_i, \qquad \overline{C} = C/N,$$

$$S = \sum_{i=1}^{N} \sin\,\theta_i, \qquad \overline{S} = S/N.$$

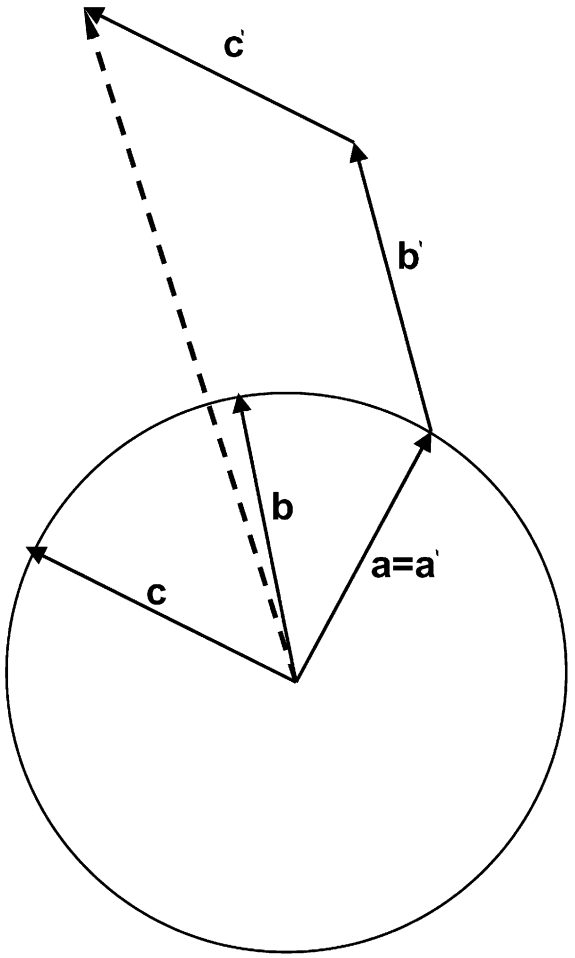

Fig. 2. Individual observed vectors (solid lines **a**, **b**, **c**) are moved parallel to themselves (**a**′, **b**′, **c**′) and combined via vector summation to form vector mean (dashed vector).

The preferred vector direction $\bar{\theta}$ is given in degrees by (cf. Fisher, 1993, p. 31)

$$\bar{\theta} = \begin{cases} \tan^{-1}(S/C), & S>0, C>0, \\ \tan^{-1}(S/C)+180, & C<0, \\ \tan^{-1}(S/C)+360, & S<0, C>0. \end{cases} \quad (1)$$

The length, R, of the combined vector may be calculated as

$$R = \left(S^2 + C^2\right)^{1/2}. \quad (2)$$

If all N of the vectors are essentially oriented in the same direction, then R will be nearly equal to N. On the other hand, if the N vectors point in all directions around the compass, then the resultant vector will be short and R near 0. Hence, a large value of R is analogous to a small variance in a linear variable. The mean resultant length is $\overline{R} = R/N$.

2.2. Von Mises and uniform distributions

Estimation of the vector mean and its length is of value, but we are interested in inference that allows us to make statements using probabilities. To make the statistical analysis rigorous, we assume distributional forms. Several distributions are appropriate, but the von Mises (also called the circular-normal) distribution is most commonly used (cf. Fisher, 1993, pp. 48–56; Mardia and Jupp, 2000, pp. 36–45). The density function for the von Mises distribution is given by

$$f(\theta) = \frac{1}{2\pi I_0(\kappa)} e^{\kappa \cos(\theta-\mu)}, \quad \begin{array}{c} 0 \le \theta < 2\pi, \\ 0 \le \mu < 2\pi, \\ \kappa \ge 0, \end{array}$$

where μ represents the preferred vector orientation of the population, and κ is the concentration parameter that indicates how closely the vectors θ cluster around μ. $I_0(\kappa)$ denotes the modified Bessel function of the first kind and order 0 (cf. Fisher, 1993, pp. 50–53; Mardia and Jupp, 2000, p. 36; Bowman, 1958, chapter 3).

The von Mises distribution is unimodal when it is wrapped around a circle; it may appear bimodal on a linear scale if the mode is near 0° or 360°. The preferred direction μ is the direction of this mode. The mode becomes broader as the concentration parameter κ decreases, and the mode disappears at $\kappa = 0$.

The uniform distribution over the circle is defined by

$$f(\theta) = \frac{1}{2\pi}, \qquad 0 \leqslant \theta < 2\pi.$$

This distribution has constant probability over the circle, so no mode exists and μ is undefined. The circular-uniform distribution corresponds to the von Mises distribution with $\kappa = 0$.

3. Script Vector_Stats

Script **Vector_Stats** operates in MATLAB version 6 on PC or UNIX machines. It requires access to the standard MATLAB functions, as well as the *Toolbox*. Operation is rapid; all simple calculations for a data set with $N = 100$ can be executed in less than 30 s on a relatively slow machine. Resampling calculations (used for small N) can take one or more minutes each.

Operation of **Vector_Stats** is similar to any MATLAB script. Begin by setting the directory path to the location of **Vector_Stats** and its component scripts and functions. In addition, make sure that the MATLAB functions and *Toolbox* are accessible. Then, in the MATLAB command window, perform any preliminary operations or calculations that are needed (e.g., load data, transform variables). Execute the script by entering **Vector_Stats** on the command line and pressing Enter.

Script **Vector_Stats** then performs some or all of the following steps:

1. Interactively requests information (data set name, dimensions, column containing data to analyze) on the data set to be processed. This can be either a text file or a MATLAB array currently in memory.
2. Interactively requests information on which calculations are to be made.
3. Calculates descriptive statistics and generates plots.
4. Tests data set for coming from uniform distribution or von Mises distribution.
5. Performs inference on vector mean, μ.
6. Performs inference on concentration parameter, κ.
7. Separates a mixture of two von Mises distributions into its components.
8. Writes calculations to external text file.

3.1. Input data

The data used for analysis may be in either of two forms: a data array already in the MATLAB memory (e.g., Table 1) or a text file (e.g., Table 2). This file begins with header records (here, three); the user instructs the script to skip these records. The data consist of the array made up of nine rows (observations) and seven columns (variables). With the exception of the header records, no character or string variables may be in the file. If such variables are in the data set, read the file separately into MATLAB and pick up the resulting array for analysis.

The script may also process a data array that already is in MATLAB's memory. If so, the user may simply specify the name of that array, and indicate which column is to be used for analysis. If the **Vector_Stats** script is used to read a text file, the data array that is input will be left in MATLAB memory, under the name **AzData**, after execution.

Table 2
Listing of file **BimodeData.txt** that contains frequency-count data for cross-bed orientations of the Salem Limestone (Sedimentation Seminar, 1966)

X-beds in Salem Ls—two-component mixture (Sedimentation Seminar, 1966)						
MidpAzim	NRegn	NLocal	NSteGen	NHarrods	Mode1	Mode2
20	51	16	9	16	51	0
60	70	59	0	1	70	0
100	43	12	3	6	43	0
140	33	7	7	9	16	17
180	47	12	4	9	0	47
220	85	38	23	10	0	85
260	85	22	2	2	0	85
300	46	20	5	4	0	46
340	28	19	4	5	14	14

Column 1 contains azimuth class midpoint and column 2 holds counts for regional observations.

The script may read individual data values, as in the array of Table 1. However, it may also read grouped data. In such cases, the azimuth values represent the midpoints of azimuth classes in one column, and a second column in the data set contains the class frequencies that are associated with the azimuth class midpoints. See the description of file **BimodeData.txt** in Section 4.1. The file is read as described above, but now two columns must be specified rather than only one.

3.2. Descriptive statistics and estimation

The script will generate descriptive statistics automatically. These include the observed vector mean, $\bar{\theta} = \hat{\mu}$, which estimates the population preferred direction μ. The maximum-likelihood estimator of μ for the von Mises distribution is the same as the general estimate given in Eq. (1). Of course, the preferred direction is not defined for a uniform distribution.

The maximum-likelihood estimator $\hat{\kappa}_{ML}$ of κ is given by the solution of

$$A_1(\hat{\kappa}_{ML}) = \overline{R}, \tag{3}$$

where $A_1(\hat{\kappa}_{ML}) = I_1(\hat{\kappa}_{ML})/I_0(\hat{\kappa}_{ML})$, I_1 and I_0 are Bessel functions, and $\overline{R} = R/N$ is the mean vector-resultant length. The function is tabled by Fisher (1993, appendix A.3, A.4) and Mardia and Jupp (2000, appendix 2.3, 2.4); the script uses their tables for estimating κ. The function A_1 and solution of Eq. (3) are discussed by Fisher (1993, pp. 50–52, 88) and Mardia and Jupp (2000, pp. 40–41).

The estimator $\hat{\kappa}_{ML}$ of Eq. (3) may be biased for small $\overline{R}$ and N. Accordingly, Fisher (1993, p. 88) provides a correction for use with $N \leqslant 15$

$$\hat{\kappa} = \begin{cases} \max\left(\hat{\kappa}_{ML} - 2/(N\hat{\kappa}_{ML}), 0\right), & \hat{\kappa}_{ML} < 2, \\ (N-1)^3\hat{\kappa}_{ML}/\left(N^3 + N\right), & \hat{\kappa}_{ML} \geqslant 2. \end{cases} \tag{4}$$

The script generates data plots (Fig. 1), as well as two Q–Q plots. A Q–Q plot compares the distributional form of the sample observations to a theoretical distribution. If the two distributions are similar, the points should fall along a straight, 45° line. Two Q–Q plots compare the observations to the circular-uniform and circular-normal distributions. These Q–Q plots, similar to those for linear data, are described by Fisher (1993, pp. 65–66, 82–83).

3.3. Testing form of population

We assume the distributional form of a population in order to use probability statements for inference, so it is necessary to determine if a sample of measurements is consistent with the assumptions. For instance, the circular-uniform distribution would imply no preferred orientation and hence no defined vector mean. On the other hand, the circular-normal distribution is widely used for inference, so it is useful to determine if the sample could have come from that distribution. This section describes three tests of the uniform distribution, and one test of circular normality.

Mardia and Jupp (2000, pp. 94–98) describe the Rayleigh test, a simple and useful way to test for uniformity. The null hypothesis is that the sample was derived from a circular-uniform distribution, versus the alternative that the distribution is not uniform. More strongly, however, "… the Rayleigh test is the most powerful invariant test against von Mises alternatives (p. 96)."

First we consider the case in which the vector mean, μ, is not known. For large samples, $2N\overline{R}^2$ is distributed as chi-square with two degrees of freedom. Mardia and Jupp (2000, p. 95) report that the modified Rayleigh statistic

$$(2N-1)\overline{R}^2 + N\overline{R}^4/2$$

also is distributed as chi-square with two degrees of freedom "… for all but the smallest sample sizes." The script uses this version of the Rayleigh test and the chi-square distribution for any selected significance level.

If the mean direction, μ, is known or hypothesized, then we are testing uniformity against an alternative having a specified mean direction. It is reasonable to consider the statistic

$$\overline{C}^* = \sum \cos(\theta_i - \mu)/N.$$

Mardia and Jupp (2000, pp. 98–99) report that $2N\overline{C}^{*2}$ is distributed as chi-square with one degree of freedom for large N. For smaller N, they provide a table (their appendix 2.6, after Stephens, 1969) of rejection points for $\overline{C}^*$; only values of significance level $\alpha = 0.10$, 0.05, 0.025, and 0.01 are tabled. The script uses the table for $N \leqslant 50$; the chi-square distribution is used for $N > 50$, without restriction on significance level α. Fisher (1993, p. 69) provides an alternative form of this test. The $\bar{R}$ and $\bar{C}^*$ tests may also be used for discrete data (Fisher, 1993, p. 71).

The third test for uniformity in **Vector_Stats** is the U-test (Mardia and Jupp, 2000, p. 104); it is based on comparison of the theoretical CDF of a uniform distribution to that of the data. This test is powerful against all alternatives, not just unimodal distributions. Sort the observed azimuths from smallest to largest over the range (0, 360), giving the order statistic $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(N)}$, where the parentheses in the subscripts represent sorted order. The theoretical uniform CDF is given by $U_i = \theta_{(i)}/2\pi$, $i = 1, \ldots, N$. Then calculate

$$U^2 = \sum_{i=1}^{N} \left[U_i - \overline{U} - \frac{i - 1/2}{N} + \frac{1}{2} \right]^2 + \frac{1}{12N}, \tag{5}$$

where $\overline{U}$ is the simple mean of the U_i. The modified statistic

$$U^{*2} = \left(U^2 - 0.1/N + 0.1/N^2\right)\left(1 + 0.8/N\right) \tag{6}$$

varies little for $N > 7$, and a table is given by Mardia and Jupp (2000, table 6.5). Only values of significance level $\alpha = 0.10$, 0.05, 0.025, and 0.01 are tabled.

Choulakian et al. (1994) provide a variant of the U-test for testing uniformity when the data are in classes (i.e., discrete distribution). The script calculates U_G^2, and tables (Choulakian et al., 1994, table 1) are used to reject the uniformity hypothesis if U_G^2 is large. Only values of significance level $\alpha = 0.10$, 0.05, 0.025, and 0.01 are used.

For circular normality, Fisher (1993, pp. 84–85) describes variants of the U-test for testing the hypothesis H_o: Data are from the von Mises distribution. This test is similar to the U-test for uniformity, except now we must take into account that μ and/or κ may be known (or hypothesized). Consider first the case where both parameters are known. This completely defines the von Mises distribution, so its CDF may be calculated. The U^{*2}-statistic is calculated by Eqs. (5) and (6); note that Fisher (1993, eq. (4.35)) contains a typographical error; see Watson (1961, eq. (26)). Reject hypothesis H_o for large U^{*2}. Fisher (1993, appendix A.8, case 0) tables U^{*2}, usable for all N and κ.

If both parameters are unknown, the CDF values U_i are calculated with the von Mises distribution, but now using the estimates $\hat{\mu}$ and $\hat{\kappa}$. Then U^2 (not U^{*2}) is calculated and tested using tabled values (cf. Fisher, 1993, appendix A.8, case 3).

Fisher (1993, pp. 84–85) discusses the other combinations of the parameters being known and unknown. Only values of significance level $\alpha = 0.10$, 0.05, 0.025, and 0.01 are used by the script, but additional values are tabled in Fisher's appendix A.8. The various U-tests can be affected by ties in the ordered data.

Fisher (1993) and Mardia and Jupp (2000) discuss other tests of uniformity and circular normality. The methods described here and written in the script were selected because they are powerful, convenient, and general.

3.4. Inference on vector mean

Analogous to ordinary linear statistics, the value of the vector mean μ is of special interest. Inference here consists of tests of hypothesis of the actual preferred direction, for the concentration parameter κ known and unknown, and confidence intervals on μ.

First consider the test of hypothesis H_0: $\mu = \mu_o$ versus the alternative H_a: $\mu \neq \mu_o$, and assume that the concentration parameter κ is known. Mardia and Jupp (2000, pp. 119–120) summarize the test. Calculate

$$\overline{C}^* = \sum \cos\left(\theta_i - \mu_o\right)/N,$$

$$\overline{S}^* = \sum \sin\left(\theta_i - \mu_o\right)/N,$$

$$\overline{R}^* = \left(\overline{C}^{*2} + \overline{S}^{*2}\right)^{1/2}/N.$$

Three test statistics are provided, depending on the value of κ and N. All are distributed as chi-square with one degree of freedom; the hypothesis is rejected for large values of the test statistic.

For large N (the script uses $N > 50$), $w = 2N\kappa\left(\overline{R}^* - \overline{C}^*\right)$ is the test statistic. For "moderate sample sizes" (undefined; the script uses $N \leqslant 50$, but applies no lower limit), two equations exist, depending on κ. If $\kappa > 2$, then

$$w^* = \left[1 - \frac{1}{4N\kappa A_1(\kappa)}\right] w,$$

where $A_1(\kappa) = I_1(\kappa)/I_0(\kappa)$. If $\kappa \leqslant 2$, then $w = 2N\gamma\left(\overline{R}^* - \overline{C}^*\right)$, where

$$\gamma = \left(\frac{1}{\kappa} + \frac{3}{8\kappa^2}\right)^{-1}.$$

Reject H_0 if w (or w^*) exceeds the tabled chi-square value with one degree of freedom and specified significance level.

Now consider the test of hypothesis H_0: $\mu = \mu_o$ versus the alternative H_a: $\mu \neq \mu_o$, assuming that the concentration parameter κ is not known. Mardia and Jupp (2000, p. 122) summarize the test, following Upton's refinements (Upton, 1973). Two test statistics are provided, both of which are appropriate for $N \geqslant 5$ but depend on the value of $\overline{C}^*$. Both statistics w are distributed as chi-square with one degree of freedom; reject the hypothesis if w is large. If $\overline{C}^* \leqslant 2/3$, then

$$w = 4N\left(\overline{R}^{*2} - \overline{C}^{*2}\right) \Big/ \left(2 - \overline{C}^{*2}\right).$$

On the other hand, if $\overline{C}^* > \frac{2}{3}$, use

$$w = \frac{2N^3}{N^2 + C^2 + 3N} \log \frac{1 - \overline{C}^{*2}}{1 - \overline{R}^{*2}}.$$

We turn now to confidence intervals about the mean preferred direction, μ. Assuming the concentration parameter κ to be unknown (Mardia and Jupp, 2000, p. 124), we can construct confidence intervals in the form $\bar{\theta} \pm \cos^{-1}(\gamma)$, where $\bar{\theta}$ is the vector mean calculated from the data. Returning to the definition of R (Eq. (2)) and $\overline{R} = R/N$, we have two choices for calculating γ, depending on $\overline{R}$. If $\overline{R} \leqslant \frac{2}{3}$, then

$$\gamma = \left[\frac{2N\left(2R^2 - N\chi^2_{1;\alpha}\right)}{R^2\left(4N - \chi^2_{1;\alpha}\right)}\right]^{1/2},$$

where $\chi^2_{1;\alpha}$ represents the $(1-\alpha)$th percentile of the chi-square distribution with one degree of freedom. If $\overline{R} > \frac{2}{3}$, then

$$\gamma = \frac{\left[N^2 - (N^2 - R^2)\exp(\chi^2_{1;\alpha}/N)\right]^{1/2}}{R}.$$

Combining $\bar{\theta}$ and $\cos^{-1}(\gamma)$, and taking the modulus 2π into account, gives the confidence interval L_1 and L_2, where $C[L_1 \leqslant \mu \leqslant L_2] = 1 - \alpha$.

Recall that L_1 and L_2 each represent azimuth directions. The confidence interval thus represents an arc centered on $\bar{\theta}$. The width of the arc is $L_2 - L_1 \pmod{2\pi}$.

For small N, some of the tests are approximate; the script calculates confidence intervals on μ using resampling (e.g., bootstrap) methods (Fisher, 1993, p. 75, 88, chapter 8). Test by determining if the hypothesized value falls within the confidence interval. Fisher (1993) and Mardia and Jupp (2000) provide other methods of inference.

3.5. Inference on concentration parameter

We are also interested in the value of the second parameter of the von Mises distribution: concentration κ of observations around the vector mean μ. Consider the test of hypothesis H_0: $\kappa = \kappa_o$ versus the alternative H_a: $\kappa \neq \kappa_o$, and assume that μ is unknown. A formal test is not described here, but instead confidence intervals about κ are defined. As usual, reject H_0 if the calculated interval does not include κ_o. Mardia and Jupp (2000, pp. 80–82, 126–127) summarize confidence intervals for various values of $\overline{R}$, the mean length of the observed resultant vector.

Case 1: $0 < \kappa < 1$, which is equivalent to $\overline{R} < 0.45$. The interval is based on a transformed variable, $g_1(2\overline{R}) = \sin^{-1}(1.2247\overline{R})$. Mardia and Jupp (2000, p. 81, eqs. (4.8.41) and (4.8.42)) state that $g_1(2\overline{R})$ follows the normal distribution, with mean approximately equal to $g_1(\kappa)$ and variance $\mathrm{var}(g_1) = 3/[4N(1-4/N)]$. They report that this approximation is satisfactory for $N \geqslant 8$. The confidence interval is calculated by first obtaining an interval based on $g_1(\kappa)$

$$L_{R,1} = \sin^{-1}(1.2247\overline{R}) - Z_{\alpha/2}\left[\mathrm{var}(g_1)\right]^{1/2},$$

$$L_{R,2} = \sin^{-1}(1.2247\overline{R}) - Z_{\alpha/2}\left[\mathrm{var}(g_1)\right]^{1/2},$$

where $Z_{\alpha/2}$ represents the $\alpha/2$ percentage point of the standard normal distribution. Then inverting function g_1 gives $L^*_{R,1} = \sin(L_{R,1})/1.2247$ and $L^*_{R,2} = \sin(L_{r,2})/1.2247$. Using $A_1(\hat{\kappa}) = \overline{R}$ (Eq. (3)) gives the $1-\alpha$ confidence limits on κ:

$$L_1 = A_1^{-1}\left(L^*_{R,1}\right), \qquad L_2 = A_1^{-1}\left(L^*_{R,2}\right).$$

For κ near 1 (i.e., $\overline{R}$ near 0.45), Mardia and Jupp (2000) recommend using Case 2. Accordingly, the script calculates both Cases 1 and 2 if $0.40 < \overline{R} < 0.45$.

Case 2: $1 < \kappa \leqslant 2$, which is equivalent to $0.45 < \overline{R} < 0.70$. This interval is also based on a transformed variable,

$$g_2(\overline{R}) = \sinh^{-1}\left(\frac{\overline{R} - c_1}{c_2}\right).$$

Mardia and Jupp (2000, p. 82, eqs. (4.8.46) and (4.8.47) state that the transformed variable is approximately normal, with mean approximately $g_2(\kappa)$ and variance $\mathrm{var}(g_2) = c_3^2/[N(1 - 3/N)]$, where $c_1 = 1.089$, $c_2 = 0.258$, and $c_3 = 0.893$. They report that this approximation is satisfactory for $N \geqslant 8$. Following a process similar to that in Case 1, we first obtain an interval based on $g_2(\kappa)$:

$$L_{R,1} = \sinh^{-1}\left[(\overline{R} - c_1)/c_2\right] - Z_{\alpha/2}\left[\mathrm{var}(g_2)\right]^{1/2},$$

$$L_{R,2} = \sinh^{-1}\left[\left(\overline{R} - c_1\right)/c_2\right] + Z_{\alpha/2}\left[\mathrm{var}(g_2)\right]^{1/2},$$

where Z represents the standard-normal percentage point. Then inverting function g_2 gives $L^*_{R,1} = c_2 \sinh(L_{R,1}) + c_1$ and $L^*_{R,2} = c_2 \sinh(L_{R,2}) + c_1$. Using $A_1(\hat{\kappa}) = \overline{R}$ (Eq. (3)) gives the $1-\alpha$ confidence limits on κ:

$$L_1 = A_1^{-1}\left(L^*_{R,1}\right), \qquad L_2 = A_1^{-1}\left(L^*_{R,2}\right).$$

Case 3: $\kappa > 2$, which is equivalent to $\overline{R} > 0.70$. Mardia and Jupp (2000, pp. 126–127, eq. (7.2.38)) give an approximate confidence interval for κ:

$$L_1 = \frac{1 + (1 + 3a)^{1/2}}{4a}, \qquad L_2 = \frac{1 + (1 + 3b)^{1/2}}{4b},$$

where $a = (N - R)/\chi^2_{N-1;1-\alpha/2}$ and $b = (N - R)/\chi^2_{N-1;\alpha/2}$; $\chi^2_{N-1;1-\alpha/2}$ and $\chi^2_{N-1;\alpha/2}$ represent the percentage points for the tails of a chi-square distribution with $N-1$ degrees of freedom. Use of the chi-square distribution here is based on a large-sample approximation; Mardia and Jupp (2000) do not offer a recommendation for its range of validity.

For small N, bootstrap resampling (Fisher, 1993, pp. 90–91, chapter 8) is used to generate a confidence interval.

3.6. Separating two components of mixture

The foregoing procedures are not generally applicable if the sample is bimodal or multimodal. For example, the vector mean calculated from a strongly bimodal distribution may point in a direction of low sample frequency. Further, if cross-bedding orientation is being studied, the two modes imply two regimes of transport conditions, and it is not reasonable to assume that a single direction can represent both adequately. We would be dealing with a mixture of distributions and should estimate the parameters of each separately.

Let us assume that we are dealing with a mixture of two von Mises distributions, which can be represented by the density function

$$f(\theta) = p\frac{1}{2\pi I_0(\kappa_1)} e^{\kappa_1 \cos(\theta - \mu_1)} + (1-p)\frac{1}{2\pi I_0(\kappa_2)} e^{\kappa_2 \cos(\theta - \mu_2)},$$

where μ_1 and κ_1 represent the vector mean and concentration of the first component, similarly for μ_2 and κ_2 of the second component, and p represents the proportion (fraction) of component 1 in the entire distribution.

Jones and James (1969) proposed using maximum-likelihood methods with numerical optimization to estimate the five parameters. However, convergence to correct estimates was found to be very sensitive to initial values of the parameters, which are difficult to obtain, and other methods subsequently have been developed. The one described here appears to work consistently.

Fisher (1993, p. 96) describes a method-of-moments estimation procedure that was proposed by Spurr and Koutbeiy (1991). This uses six estimating equations for the five parameters, but seems to work better than methods using five equations; these six equations are

$$\begin{aligned}
&pA_1(\kappa_1)\cos(\mu_1) + (1-p)A_1(\kappa_2)\cos(\mu_2) = \overline{C_1},\\
&pA_2(\kappa_1)\cos(2\mu_1) + (1-p)A_2(\kappa_2)\cos(2\mu_2) = \overline{C_2},\\
&pA_3(\kappa_1)\cos(3\mu_1) + (1-p)A_3(\kappa_2)\cos(3\mu_2) = \overline{C_3},\\
&pA_1(\kappa_1)\sin(\mu_1) + (1-p)A_1(\kappa_2)\sin(\mu_2) = \overline{S_1},\\
&pA_2(\kappa_1)\sin(2\mu_1) + (1-p)A_2(\kappa_2)\sin(2\mu_2) = \overline{S_2},\\
&pA_3(\kappa_1)\sin(3\mu_1) + (1-p)A_3(\kappa_2)\sin(3\mu_2) = \overline{S_3},
\end{aligned}$$

where

$$A_r(\kappa) = I_r(\kappa)/I_0(\kappa), \tag{7}$$

$$\overline{C_r} = \sum \cos(r\theta_i)/N,$$

$$\overline{S_r} = \sum \sin(r\theta_i)/N$$

for $r = 1, 2, 3$. Then let $\tilde{\mu}_1$, $\tilde{\kappa}_1$, $\tilde{\mu}_2$, $\tilde{\kappa}_2$, and $\tilde{p}$ be any set of estimates of the parameters, and consider ΔC_1, ΔC_2, ΔC_3, ΔS_1, ΔS_2, ΔS_3, to be deviations of the estimated model from the observed statistics; for example, the first of the six equations gives

$$\Delta C_1 = \tilde{p}A_1(\tilde{\kappa}_1)\cos(\tilde{\mu}_1) + (1-\tilde{p})A_1(\tilde{\kappa}_2)\cos(\tilde{\mu}_2) - \overline{C_1}$$

and similarly for the other five. Then the sum-of-squares criterion is defined as

$$\Lambda^2(\tilde{\mu}_1, \tilde{\mu}_2, \tilde{\kappa}_1, \tilde{\kappa}_2, \tilde{p}) = \Delta C_1^2 + \Delta C_2^2 + \Delta C_3^2 + \Delta S_1^2 + \Delta S_2^2 + \Delta S_3^2.$$

MATLAB function lsqnonlin (in the *Optimization Toolbox*) is used by the script to minimize Λ^2 and estimate the five parameters. The function requires the user to specify a tolerance for convergence on the five parameters.

This method may be extended to more than two components. However, the computational requirements grow substantially. It should be noted that obtaining good estimates of two components can be difficult, and the complexity grows for more modes if no obvious initial estimates exist. In general, separating components of a distribution is difficult with azimuthal data unless the modes are very clearly defined and the sample size is large.

Fisher (1993, pp. 99–100) and Mardia and Jupp (2000, p. 91) discuss special cases that may help establish initial estimates. A common occurrence is that directions of the two modes differ from each other by 180°. If we assume that these two mean directions are μ and $\mu + 180$ and concentrations are $\kappa_1 = \kappa_2 = \kappa$, then a simple calculation allows us to estimate the resulting three

parameters. Fisher proposes calculating

$$\psi_i = \theta_i \, [mod\, 180^\circ], \qquad i = 1, 2, \ldots, N,$$

$$C_\psi = \sum \cos(2\psi_i), \qquad S_\psi = \sum \sin(2\psi_i),$$

$$\overline{R}_\psi = \left(C_\psi^2 + S_\psi^2\right)^{1/2}/N.$$

Estimate μ_ψ from Eq. (1) using C_ψ and S_ψ. Then the three parameters can be found

$$\hat{\mu} = \hat{\mu}_\psi/2,$$

$$\hat{\kappa}^* = A_2^{-1}\left(\overline{R}_\psi\right),$$

$$(2\hat{p} - 1)A_1(\hat{\kappa}^*) = \overline{C_1}\,\cos(\hat{\mu}) + \overline{S_1}\,\sin(\hat{\mu}),$$

where the terms are defined in Eq. (7). Eq. (4) is used to convert $\hat{\kappa}^*$ to $\hat{\kappa}$. The script automatically calculates these values.

Fig. 3 shows a rose diagram of the data in Table 2 (column 1, midpoints of 40° classes; column 2, frequency for regional observations). This data set was obtained by Sedimentation Seminar (1966), and consists of 488 cross-bed observations in the middle Mississippian Salem Limestone of central Indiana. The rose diagram shows clear bimodality, so we estimate the two components separately. We do not formally use the three-parameter short cut to get initial estimates, but will get them by eye. It appears that the mean direction of the larger of the two modes is about $\tilde{\mu}_1 = 240^\circ$, and the other vector mean is about $\tilde{\mu}_1 = 60^\circ$. Although mode two seems more concentrated than the other, we will use $\tilde{\kappa}_1 = \tilde{\kappa}_2 = 2$. The fraction of mode one appears to be about 0.65 of the observations.

Application of the script (see data set **BimodeDataResults.txt**) gives estimates of the two components: $\hat{\mu}_1 = 238.8$, $\hat{\kappa}_1 = 1.48$, $\hat{\mu}_2 = 57.2$, $\hat{\kappa}_2 = 1.89$, and $\hat{p} = 0.62$. Note that the two estimated mean directions differ by about 180°. The authors interpreted the bimodality as arising from "… oscillating tidal currents on a shallow marine shelf …" (Sedimentation Seminar, 1966, p. 95), which is consistent with the estimates.

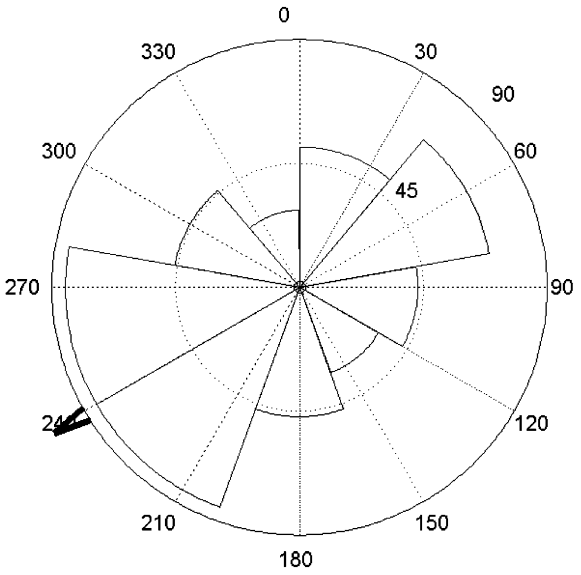

Fig. 3. Rose diagram showing bimodally distributed cross-beds in the Salem Limestone (Sedimentation Seminar, 1966).

4. Other information

4.1. Material in IAMG server

The IAMG server contains the following:

- The script **Vector_Stats**, which consists of 33 MATLAB m-files (scripts and functions). These m-files may be modified by users for their own needs.
- A file titled **READ_ME_STATS.txt** that describes operation of the script.
- A file titled **TestData.txt** that contains cross-bed azimuths from the Bulgoo Formation in the Belford Anticline, Australia. This data is Set 2 in Fisher (1993, appendix B.6); see also Fisher and Powell (1989). The data array is shown in Table 1.
- A file titled **TestDataResults.txt** TestDataResults.txt that contains the text file generated by analysis of **TestData.txt**.
- A file titled **BimodeData.txt** that contains azimuths of cross-beds in the Salem Limestone (Sedimentation Seminar, 1966). The data set is shown in Table 2 and Fig. 3. This sample is strongly bimodal, and may be used for separating components of a mixture. Note that this data set contains grouped data. Column 1 contains the azimuth-class midpoints and column 2 contains counts of the observations in each azimuth class (their regional observations). For comparison, note that the sixth and seventh columns contain the numbers of entries corresponding to each mode.
- A file titled **BimodeDataResults.txt** that contains the text file generated by analysis of the regional observations in **BimodeData.txt**.

4.2. Axial variables

This paper and the script are aimed at vectorial data, that is, each measurement represents a single-headed vector direction. Direction of cross-bedding is a typical example of this sort of data. However, geological measurements commonly include orientations of axial variables. Axial variables represent undirected lines;

these phenomena arise from processes that leave evidence of orientation but do not allow distinguishing between direction θ and $\theta+180^\circ$. Examples of these include groove marks on the soles of turbidite beds, current lineations in sandstones, orientations of long axes of pebbles in glacial till, and orientations of feldspar laths.

A geologist developed the procedure for analyzing axial data (Krumbein, 1939). Following Krumbein, Fisher (1993, p. 37) recommends the following process. Firstly, convert the axial directions ψ to vectors θ by $\theta_i = 2\psi_i \,[\mathrm{mod}\, 360]$ for all i. Secondly, analyze the resulting vectors using the methods described in this paper. Finally, convert the estimated preferred direction $\hat{\theta}$ back to the axial direction by dividing by 2. Similarly halve the width of the confidence interval on the preferred direction. Fisher recommends leaving the measures of concentration in vectorial units.

Acknowledgments

Thanks to John Davis and an anonymous reviewer for helpful suggestions and information. John also reminded me that Bill Krumbein made an early contribution to analysis of orientation data.

References

Bowman, F., 1958. Introduction to Bessel Functions. Dover Publications, New York, 135pp.

Blaesild, P., Granfeldt, G., 2003. Statistics with Applications in Biology and Geology. Chapman & Hall/CRC, Boca Raton, FL, 555p.

Choulakian, V., Lockhart, R.A., Stephens, M.A., 1994. Cramer–von Mises statistics for discrete distributions. Canadian Journal of Statistics 22 (1), 125–137.

Fisher, N.I., 1993. Statistical Analysis of Circular Data. Cambridge University Press, New York, 277pp.

Fisher, N.I., Powell, C.McA., 1989. Statistical analysis of two-dimensional palaeocurrent data: Methods and examples. Australian Journal of Earth Science 36 (1), 91–107.

Hanselman, D., Littlefield, B., 2001. Mastering MATLAB 6: A Comprehensive Tutorial and Reference. Prentice-Hall, Upper Saddle River, NJ, 814pp.

Hunt, B.R., Lipsman, R.L., Rosenberg, J.M., 2001. A Guide to MATLAB for Beginners and Experienced Users. Cambridge University Press, Cambridge, 327pp.

Jones, T.A. MATLAB functions to analyze directional (azimuthal) data, II: correlation. Computers & Geosciences 32, in press.

Jones, T.A., James, W.R., 1969. Analysis of bimodal orientation data. Mathematical Geology 1 (2), 129–135.

Krumbein, W.C., 1939. Preferred orientation of pebbles in sedimentary deposits. Journal of Geology 47, 673–706.

Mardia, K.V., 1972. Statistics of Directional Data. Academic Press, London, 357pp.

Mardia, K.V., Jupp, P.E., 2000. Directional Statistics. Wiley, Chichester, 429pp.

Middleton, G.V., 2000. Data Analysis in the Earth Sciences Using MATLAB. Prentice-Hall, Upper Saddle River, NJ, 260p.

Sedimentation Seminar, 1966. Cross-bedding in the Salem Limestone of Central Indiana. Sedimentology 6 (1), 95–114.

Spurr, B.D., Koutbeiy, M.A., 1991. A comparison of various methods for estimating the parameters in mixtures of von Mises distributions. Communications Statistical-Simulation Computing 20, 725–741.

Stephens, M.A., 1969. Tests for randomness of directions against two circular alternatives. Journal of American Statistical Association 64 (2), 280–289.

Upton, G.J.G., 1973. Single-sample tests for the von Mises distribution. Biometrika 60 (1), 87–99.

Watson, G.S., 1961. Goodness-of-fit tests on a circle. Biometrika 48 (1), 109–114.

Wells, N.A., 1999. ASTRA.BAS: a program in QuickBasic 4.5 for exploring rose diagrams, circular histograms, and some alternatives. Computers & Geosciences 25 (6), 641–654.