ELSEVIER

Computers & Geosciences 32 (2006) 176–183

COMPUTERS & GEOSCIENCES

www.elsevier.com/locate/cageo

MATLAB functions to analyze directional (azimuthal) data—II: Correlation☆

Thomas A. Jones*

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

Received 28 November 2004; received in revised form 28 June 2005; accepted 30 June 2005

Abstract

Data that represent azimuthal directions cannot be analyzed in the same manner as ordinary linear variables that measure, for example, size or quality. This is especially true for correlations between variables. Although many statistical techniques have been developed for dealing with directional data, software is scarce for the unique calculations. This paper describes a MATLAB script, located on the IAMG server, that calculates measures of association (correlations) between two circular variables, or between a circular and a linear variable. The calculations include tests of significance of the associations. These tests use large-sample approximations or resampling methods.
© 2005 Elsevier Ltd. All rights reserved.

Keywords: Association; Azimuths; Circular–circular; Circular–linear; Linear–circular; Vector_Corr

1. Introduction

Data that describe directions, such as the azimuthal directions of cross-beds, are important in the earth sciences. However, observations from such directional variables cannot be analyzed with standard statistical methods because they are cyclical in nature, modulo 360°. Statisticians and earth scientists have developed special ways to statistically analyze azimuthal or other cyclical data. Excellent references on cyclical variables include Mardia (1972), Fisher (1993), and Mardia and Jupp (2000).

General software for analyzing azimuthal data is not widely available, but code includes MATLAB scripts (Middleton, 2000), SAS macros (Blaesild and Granfeldt, 2003), Stata (CIRCSTAT by N.J. Cox, 1998, on http://fmwww.bc.edu/repec/bocode/c/circstat.zip), and the DDSTAP package developed by Ashis SenGupta (Indian Statistical Institute, Calcutta). Software to form a general system is provided by Jones (2005) and this paper, as well as other scripts under development.

Although extensive methods were developed initially for analyzing a single variable measured in a sample, tools for studying multiple-variable samples have been developed more recently. Such methods are related to correlation and regression. Correlation between two linear variables, commonly in the context of the bivariate normal distribution, is well known. However, similar measures of association between two circular variables, or between a circular and a linear variable, are not as well known.

MATLAB® (a registered trademark of The Mathworks, Inc.) is a general, widely used program for doing mathematical computation. Many good references are available, including Hanselman and Littlefield (2001) and Hunt et al. (2001). Basic functionality in MATLAB

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

*Corresponding author. 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.2005.06.021

can be combined into scripts to execute complex statistical analyses.

This paper describes a MATLAB script that is designed for analysis of relations between azimuthal and linear data. The script, named **Vector_Corr**, calculates correlation coefficients and significance tests between pairs of circular–circular and linear–circular variables. The script follows the same operational methods and data input as script **Vector_Stats**, described by Jones (2005).

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

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 at which an event (e.g., tornado) may occur is cyclical over 24 h, and day of the year is cyclical. Converting such cyclical variables to pseudo-angles (say, azimuths) allows us to analyze these variables.

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

2. Overview of orientation data

Assume we have taken measurements of cross-bed orientations, $\boldsymbol{\theta}_i$, that represent a direction at each of N locations. The $\boldsymbol{\theta}_i$ are in the form of azimuths, with 0° representing North, 90° East, 180° South, and so on. Hence, the values of the $\boldsymbol{\theta}_i$ are in the range (0°, 360°), or equivalently (0, 2π) radians. This definition of azimuth is that used by geologists. Different definitions are used in mathematics and some other disciplines. We may think of each $\boldsymbol{\theta}_i$ as a unit vector that points in the direction of interest. Assume that we have measured other variables, perhaps linear (e.g., size and quantity) and circular (azimuthal), at the same locations.

Table 1 shows a listing of a data array such as used by the script. Each row in the array represents an observation. The columns represent two variables, one of which is azimuthal and the other is linear. Column 1 contains concentrations of ozone at a measuring station, and column 2 contains the wind direction at the time of measurement. These data are provided on the IAMG server in file **WindOzoneData.txt**, and are discussed in Section 4.1.

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

28.0	327
85.2	91
80.5	88
4.7	305
45.91	344
12.7	270
72.52	67
56.62	21
31.5	281
112.	8
20.0	204
72.51	86
16.0	333
45.92	18
32.6	57
56.61	6
52.6	11
91.8	27
55.2	84

Column 1 contains ozone concentration and column 2 contains wind direction (Fisher, 1993, appendix B.18).

Fisher (1993, Chapter 6) shows several ways to display two related variables, both circular–circular and linear–circular. Fig. 1 shows a cross plot of a linear variable (ozone concentration) along the horizontal axis versus an azimuthal variable (wind direction) on the vertical axis. Note that the vertical axis extends from 0° to 720°. In order to show the distribution of directions without a split at the 0–360 boundary, the original azimuth values ranging over (0°, 360°) are repeated over the range (360°, 720°). The horizontal line at 360° shows where the azimuths repeat. Here the cluster of points along this line (ranging from about 200°–450°) is easier to interpret than the two groups (0°–100° and 200°–360°). For instance, ozone concentration is lower when the wind direction is northwesterly (270–360), and higher for a northeasterly direction (0–90).

For two azimuthal variables, both axes are extended to 720°. Fig. 2 shows a cross plot of two azimuthal variables from a second data set (**FluvialReachXbedData.txt**; see Section 4.1). The observations were taken in the Cretaceous Rocktown channel sandstone of Kansas. One variable is orientation of relatively straight reaches in the latest stages of fluvial deposition of the river system, and the other is the vector-mean orientation of cross beds in each reach. Fig. 3 shows a cross plot from **PairedWindData.txt** (see Section 4.1). Here lineations oriented approximately at 45° are easier to see in the central portions of the plot than having to use

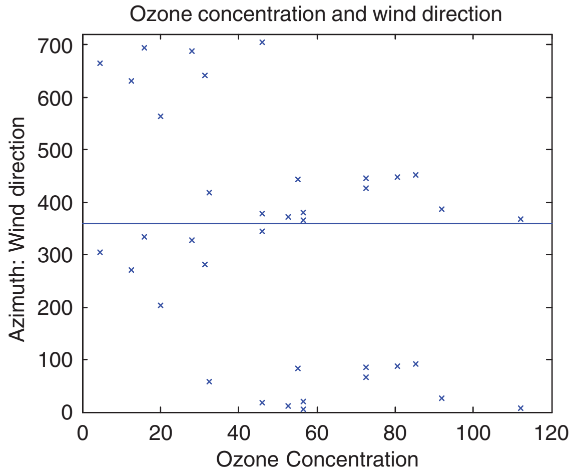

Fig. 1. Cross plot of linear variable (ozone concentration) versus azimuthal variable (wind direction), showing relation between variables.

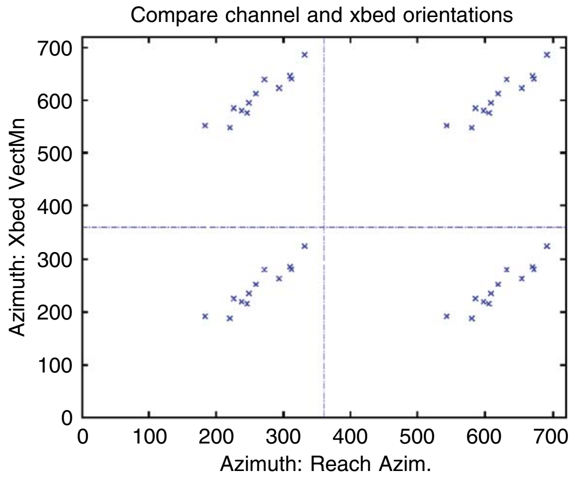

Fig. 2. Cross plot of showing relation between two azimuthal variables: orientations of straight reaches in a paleo-river versus vector means of measured cross beds in these reaches.

the subgroups located in the edges and corners of the plot.

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). A useful estimate of the vector mean (direction of preferred orientation, $\bar{\theta}$) is given by the vector sum. Define $C = \sum_{i=1}^{N} \cos\theta_i$ and $S = \sum_{i=1}^{N} \sin\theta_i$, where the θ_i represent the N observed directions. The preferred vector direction

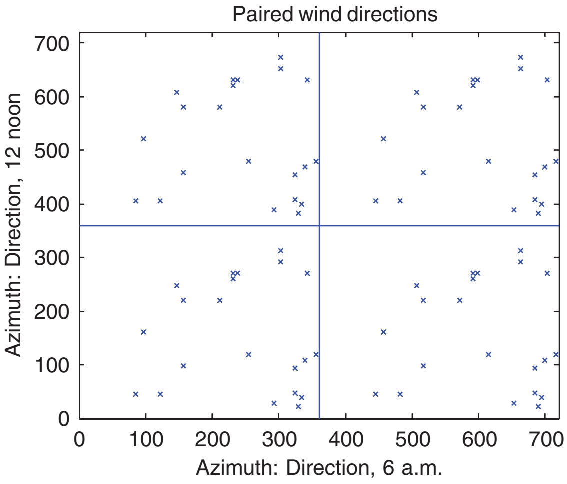

Fig. 3. Cross plot of two azimuthal variables: wind direction at 6 a.m. versus wind direction at 12 noon.

$\bar{\theta}$ is given in degrees by

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

(cf. Fisher, 1993, p. 31; Jones, 2005, Eq. (1)).

The length, R, of the combined vector may be calculated as $R = (S^2 + C^2)^{1/2}$.

If all N of the vectors are oriented essentially in the same direction, then R should have length nearly equal to N, analogous to small variance in a linear variable. The mean resultant length is

$$\bar{R} = R/N. \qquad (2)$$

The von Mises (also called the circular–normal) distribution is widely used as a distributional form for azimuths (cf. Fisher, 1993, p. 48–56; Mardia and Jupp, 2000, p. 36–45). Its parameters consist of $\boldsymbol{\mu}$ (the preferred vector orientation of the population), and $\boldsymbol{\kappa}$ (the concentration parameter that indicates how closely the data vectors $\boldsymbol{\theta}$ cluster around $\boldsymbol{\mu}$). The sample vector mean, $\bar{\boldsymbol{\theta}}$, estimates the population preferred direction $\boldsymbol{\mu}$. The estimator of $\boldsymbol{\kappa}$ is a function of $\bar{\boldsymbol{R}}$, and is usually determined using tables. The uniform distribution over the circle has constant probability for all directions, similarly to the linear uniform distribution, and is an optional distribution. With constant probability, no mode exists and $\boldsymbol{\mu}$ is undefined.

3. Script `Vector_Corr`

Script **`Vector_Corr`** operates in MATLAB version 6 on the PC or UNIX machine. It requires access to the standard MATLAB functions, as well as the *MATLAB Toolbox*. Operation is very rapid; calculations for a data set with six variables and $N = 100$ can be executed in less than 20 s on a relatively slow machine.

Operation of **Vector_Corr** is similar to **Vector_Stats** and other MATLAB scripts. Begin by setting the directory path to the location of **Vector_Corr** 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_Corr** on the command line and pressing Enter.

Script **Vector_Corr** performs some or all of the following steps, depending on commands:

1. Interactively requests information (dataset name, dimensions, which columns in the array contain data to be analyzed) 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 circularly based cross plots.
4. For each pair of circular–circular or linear–circular variables to be analyzed, it calculates one or more correlations (depending on various assumptions and hypotheses) and performs tests of significance.
5. 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 MATLAB memory (e.g., Table 1) or a text file (e.g., **FluvialReachXbedData.txt**; Table 2). The text file begins with header records (here, three); the user instructs the script to skip these records. This data set follows with an array made up of 12 rows (observations) and five columns (variables). Columns 3 and 4 were used in Fig. 2. 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. See Jones (2005) for a further description of input files.

If the data array already is in MATLAB's memory, the user may simply specify the name of that array, and indicate which columns are to be used for analysis. If the **Vector_Stats** or **Vector_Corr** 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.

Vector_Corr does not process grouped data.

Table 2
Listing of typical data set used by **Vector_Corr**

Rocktown channel SS—reach trends and Xbed azimuths
Siemers, 1976, J. Sed. Petr., p. 97–123—Table 4

Seg	Length	ReachAzim	XbedVectMn	N
1	6.0	312	280	34
2	3.5	238	220	78
3	3.5	310	285	51
4	1.0	249	235	12
5	2.5	183	192	10
6	0.5	271	279	13
7	4.5	332	325	66
8	2.0	226	225	10
10	4.0	246	216	59
11	1.5	259	252	13
12	2.5	220	188	48
13	3.5	294	263	26

Column 3 contains orientations of straight reaches in a paleoriver and column 4 contains vector mean of cross bed measurements within these reaches.

3.2. Types of correlations

Fisher (1993, Section 6.1) introduces the analysis of two correlated variables in terms of regression. Specifically, he reminds us of the common use of a straight-line relationship between a variable X and the mean of another variable Y, as $E[Y|X = x] = a + bx$. A use of correlation is to quantify the degree to which such a relation explains the variation in Y. This tie to regression is the main impetus for his discussion of correlation between two circular variables, or between a linear and a circular variable.

The association between two circular random variables, say $\mathbf{\Theta}$ and $\mathbf{\Phi}$, is referred to as circular–circular association. The association between linear and circular variables can be of two types. Linear–circular association is defined to be the situation of predicting the mean value of X given $\boldsymbol{\Theta} = \theta$ On the other hand, circular–linear association is concerned with predicting the vector mean of $\mathbf{\Theta}$ given $X = x$. The measures of association for these three cases are different.

Fisher (1993) and Mardia and Jupp (2000) provide several measures for each of these types of associations. These are based on various models, containing few to more complex assumptions. It should be noted that these various measures do not estimate the same thing, so the value obtained from one coefficient cannot be compared to the value from a second coefficient. Fisher (1993) and Mardia and Jupp (2000) also provide tests of significance for the measures, most of which require large sample sizes.

Fisher suggests that resampling methods be used for some tests, especially for small samples. The **Vector_Corr** script uses resampling if $N \leqslant 30$–35. For comparison purposes, the standard methods are also provided, even for smaller N. The resampling method is summarized in Section 3.5.

3.3. Circular–circular correlation

We first consider the association between two circular random variables, $\mathbf{\Theta}$ and $\mathbf{\Phi}$. Assume we have N pairs of azimuthal observations, $(\boldsymbol{\theta}_1, \boldsymbol{\phi}_1)$, $(\boldsymbol{\theta}_2, \boldsymbol{\phi}_2)$, …, $(\boldsymbol{\theta}_N, \boldsymbol{\phi}_N)$. There are several choices for a measure of association between the variables.

3.3.1. Embedding: canonical correlation

Mardia and Jupp (2000, p. 248–249) present the "embedding approach." This splits each of the azimuths into its sine and cosine components. Hence, we have four variables $(\sin\boldsymbol{\theta}, \cos\boldsymbol{\theta}, \sin\boldsymbol{\phi}, \cos\boldsymbol{\phi})$ and canonical correlation coefficients for this group can be calculated. Jupp and Mardia (1980) define $r_{cc} = \text{corr}(\cos\boldsymbol{\theta}, \cos\boldsymbol{\phi})$, $r_{sc} = \text{corr}(\sin\boldsymbol{\theta}, \cos\boldsymbol{\phi})$, and similarly for r_{cs} and r_{ss}, where corr$(u,\ v)$ represents the usual correlation coefficient between two linear variables u and v. They also define $r_1 = \text{corr}(\cos\boldsymbol{\theta}, \sin\boldsymbol{\theta})$ and $r_2 = \text{corr}(\cos\boldsymbol{\phi}, \sin\boldsymbol{\phi})$. Then the sum of the squared canonical correlation coefficients is given by

$$r^2 = [A + B + C + D]/[(1 - r_1^2)(1 - r_2^2)],$$

where

$$\begin{aligned} A &= r_{cc}^2 + r_{cs}^2 + r_{sc}^2 + r_{ss}^2, \\ B &= 2(r_{cc}r_{ss} + r_{cs}r_{sc})r_1r_2, \\ C &= -2(r_{cc}r_{cs} + r_{sc}r_{ss})r_2, \\ D &= -2(r_{cc}r_{sc} + r_{cs}r_{ss})r_1. \end{aligned}$$

Thanks to P.E. Jupp and K.V. Mardia for providing the correction to a typographical error in their book; future printings will have corrected this. The example calculations (p. 249) also contain several typographical errors.

Independence of the variables is implied by r^2 near 0, so the association is significant for large r^2. Under the hypothesis H_o of independence and for large N, Nr^2 is distributed approximately as χ^2 with four degrees of freedom; reject H_o for large Nr^2. This test is calculated by the script if $N \geqslant 15$; a resampling test using values of r^2 is conducted for $N \leqslant 30$.

3.3.2. T-monotone association

If the circular ranks u_i of the $\boldsymbol{\phi}_i$ are the same as ranks v_i of the $\boldsymbol{\theta}_i$, then $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are said to be concordant; on the other hand, if the ranks of $\boldsymbol{\phi}$ and $\boldsymbol{\theta}$ are reverse-ordered, then they are discordant. T-monotone association measures the degree of concordance, discordance, or non-association. Fisher (1993, p. 148–149) and Mardia and Jupp (2000, p. 252) describe a measure, $\hat{\prod}_N$, that equals 1 for perfect concordance and -1 for perfect discordance. This measure is based on the circular ranks of $\boldsymbol{\theta}$ and $\boldsymbol{\phi}$ (u_i and v_i, respectively) for $i = 1, 2, \ldots, N$. These ranks are converted to uniform scores, $\beta = 2\pi u_i/N$ and $\gamma_i = 2\pi v_i/N$. Then

$$\hat{\prod}_N = 4(AB - CD)/N^2,$$

where

$$\begin{aligned} A &= \sum \cos\beta_i \cos\gamma_i, \\ B &= \sum \sin\beta_i \sin\gamma_i, \\ C &= \sum \cos\beta_i \sin\gamma_i, \\ D &= \sum \sin\beta_i \cos\gamma_i. \end{aligned}$$

Printed values from the script are labeled PIhatN. Non-association is rejected for large values of $|\hat{\prod}_N|$; Fisher (1993, Appendix A.13) provides a table of the distribution of $(N-1)|\hat{\prod}_N|$ for $N \geqslant 8$. If $N \leqslant 30$, the resampling test is also calculated.

The estimate of $\hat{\prod}_N$ is affected by tied values in the ranks of $\boldsymbol{\theta}_i$ and $\boldsymbol{\phi}_i$. The test is also affected. The extent of this effect may be tested by slightly modifying the tied values over several trials and looking at how $\hat{\prod}_N$ varies.

3.3.3. T-linear association

T-linear association is based on the concept that $\mathbf{\Theta}$ and $\mathbf{\Phi}$ are related by model M1: $\mathbf{\Phi} = \mathbf{\Theta} + \boldsymbol{\theta}_\text{o}$ (modulo 360°) or M2: $\mathbf{\Phi} = -\mathbf{\Theta} + \boldsymbol{\theta}_\text{o}$ (modulo 360°). This is the strongest form of dependence between the two random variables. Fisher (1993, p. 151–153) and Mardia and Jupp (2000, p. 249–250) discuss a measure of T-linear association,

$$\hat{\rho}_T = 4(AB - CD)/[(N^2 - E^2 - F^2)(N^2 - G^2 - H^2)]^{1/2},$$

where

$$\begin{aligned} A &= \sum \cos\theta_i \cos\varphi_i, & B &= \sum \sin\theta_i \sin\varphi_i, \\ C &= \sum \cos\theta_i \sin\varphi_i, & D &= \sum \sin\theta_i \cos\varphi_i, \\ E &= \sum \cos 2\theta_i, & F &= \sum \sin 2\theta_i, \\ G &= \sum \cos 2\phi_i & H &= \sum \sin 2\varphi_i. \end{aligned}$$

If $\hat{\rho}_T$ is near 1, model M1 is more likely true, and a value near -1 implies model M2. Independence thus is rejected for large $|\hat{\rho}_T|$. The measure is labeled in the script's printout as rhoThat.

The distribution of $\hat{\rho}_T$ is not known for small N. The resampling test is conducted for $N \leqslant 35$. For $N \geqslant 25$, there are two possible ways to calculate a critical test value, depending on the marginal distributions of $\mathbf{\Theta}$ and $\mathbf{\Phi}$ (Fisher, 1993, p. 152):

1. If either $\mathbf{\Theta}$ or $\mathbf{\Phi}$ has mean vector length $\bar{R} = R/N$ near zero, then $N\hat{\rho}_T$ approximately follows the double-exponential distribution with density $f(x) = 0.5\exp(-|x|)$. Hence, for significance level $\boldsymbol{\alpha}$, reject the two-sided null hypothesis of no T-linear

association ($\mathrm{H_o}: \rho_T = 0$) if $|N\hat{\rho}_T| > -\log_e(\boldsymbol{\alpha})$. For a one-sided hypothesis, use $-\log_e(2\boldsymbol{\alpha})$.

2. If we cannot assume the mean vector length is near zero for either variable, then a normal approximation is available. Calculate the vector mean directions ($\bar{\theta}$ and $\bar{\phi}$) and the mean vector lengths ($\bar{R}_\theta = R_\theta/N$ and $\bar{R}_\phi = R_\phi/N$) (Eqs. (1) and (2)), and then

$$\hat{\alpha}_\theta = (1/N)\sum \cos 2(\theta_i - \bar{\theta}),$$
$$\hat{\beta}_\theta = (1/N)\sum \sin 2(\theta_i - \bar{\theta}),$$
$$\hat{\alpha}_\phi = (1/N)\sum \cos 2(\phi_i - \bar{\phi}_i),$$
$$\hat{\beta}_\phi = (1/N)\sum \sin 2(\phi_i - \bar{\phi}_i),$$

followed by

$$U_\theta = \left(1 - \hat{\alpha}_\theta^2 - \hat{\beta}_\theta^2\right)/2,$$
$$U_\phi = \left(1 - \hat{\alpha}_\phi^2 - \hat{\beta}_\phi^2\right)/2,$$
$$V_\theta = \bar{R}_\theta^2(1 - \hat{\alpha}_\theta),$$
$$V_\phi = \bar{R}_\phi^2(1 - \hat{\alpha}_\phi).$$

Then the variable

$$Z = U_\theta U_\phi \hat{\rho}_T [N/(V_\theta V_\phi)]^{1/2}$$

is distributed approximately normally, with mean 0 and variance 1. Hence, reject the two-sided hypothesis $Ho: \rho_T = 0$ if $|Z| > Z_{\alpha/2}$, where $Z_{\alpha/2}$ is taken from the normal table. An obvious modification applies to one-sided tests.

3.4. Linear–circular correlation

Now consider the relation between an azimuthal and a linear variable, $\boldsymbol{\Theta}$ and X. We have N pairs of observations (x_1, θ_1), …, (x_N, θ_N). Several measures are available.

First we briefly consider circular–linear association in which we are interested in predicting the mean of $\boldsymbol{\Theta}$ for a given value of X. This can be changed into a circular–circular situation (Fisher, 1993, p. 160–161) by transforming the X values to angles, as $\boldsymbol{\Phi} = 2\tan^{-1}(X)$. With the problem changed to two circular variables, test for significant association with the T-linear ($\hat{\rho}_T$) method discussed in Section 3.3. We will not discuss the circular–linear case further.

3.4.1. Embedding: C-linear association

This linear–circular model divides the azimuthal variable into two components—sine and cosine—for each observation. It then uses the sample multiple correlation coefficient $R_{x\theta}$ of X with ($\cos\boldsymbol{\Theta}$, $\sin\boldsymbol{\Theta}$) as a measure of association. Mardia and Jupp (2000, p. 245–246) and Fisher (1993, p. 145) state that

$$R_{x\theta}^2 = \frac{r_{xc}^2 + r_{xs}^2 - 2r_{xc}r_{xs}r_{cs}}{1 - r_{cs}^2}$$

where $r_{xc} = \mathrm{corr}(x, \cos\boldsymbol{\theta})$, $r_{xs} = \mathrm{corr}(x, \sin\boldsymbol{\theta})$, and $r_{cs} = \mathrm{corr}(\cos\boldsymbol{\theta}, \sin\boldsymbol{\theta})$ are ordinary linear correlation coefficients.

Reject the hypothesis of no association ($H_0: \rho_{x\theta}^2 = 0$) if $R_{x\theta}^2$ is large. Mardia and Jupp (2000, p. 246) state that if X is distributed normally, then $(N-3)R_{x\theta}^2/(1 - R_{x\theta}^2)$ is distributed under $\mathrm{H_o}$ as F with 2 and N−3 degrees of freedom. Rejection suggests that the associated regression model can be fit. Liddell and Ord (1978) give an exact distribution for the special case of $\rho_{x\theta}^2 > 0$. The F-test is calculated if $N \geqslant 20$. Resampling is used if $N \leqslant 35$, without an assumption of distributional form.

3.4.2. C-association

This test extends the $R_{x\theta}^2$ estimator to use with ranks. As with the T-monotonic association ($\hat{\prod}_N$), define the circular ranks of the azimuths $\boldsymbol{\theta}_i$ to be u_i and convert to uniform scores by $\boldsymbol{\beta}_i = 2\pi u_i/N$. Let the usual linear ranks of the x_i be defined as v_i. Fisher (1993, p. 140–141) and Mardia and Jupp (2000, p. 246–248) show that

$$D_N = a_N(T_c^2 + T_s^2)$$

is a measure of C-association, where

$$T_c = \sum v_i \cos\beta_i, \qquad T_s = \sum v_i \sin\beta_i$$

and

$$a_N = \begin{cases} 1/[1 + 5\cot^2(\pi/N) + 4\cot^4(\pi/N)] & N \text{ even,} \\ 2\sin^4(\pi/N)/[1 + \cos(\pi/N)]^3 & N \text{ odd.} \end{cases}$$

D_N is defined over (0, 1), with values near 0 indicating lack of C-association.

For testing, calculate

$$U_N = 24(T_c^2 + T_s^2)/(N^3 + N).$$

Reject the hypothesis that X and $\boldsymbol{\Theta}$ are independent if U_N is large. The script uses a tabled distribution of U_N for $6 \leqslant N < 100$; U_N is distributed as χ^2 with two degrees of freedom for larger N. A resampling test is calculated if $N \leqslant 30$.

D_N and U_N are affected by tied values in the ranks of $\boldsymbol{\theta}_i$ and x_i. The extent of this effect may be tested by slightly modifying the tied values over several trials and looking at how D_N and U_N vary.

3.5. Resampling

A common problem for inference is that the exact sampling distributions of many test statistics are unknown. In many situations, large-sample approximations to the distributions are used to allow testing. However, these approximations may be poor for small-to-moderate sample sizes. Resampling methods have been developed to deal with such problems. These are computationally intensive, but modern computers allow resampling to be done almost routinely.

Fisher (1993, Chapter 8) discusses two types of resampling: bootstrap and randomization methods. Bootstrap methods are used by script **`Vector_Stats`** (Jones, 2005) for single-sample inference, including confidence intervals. Randomization tests (also called permutation tests) are used in the **`Vector_Corr`** script for testing hypotheses of no association or correlation.

Assume we have two azimuthal and/or linear variables, $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$, with N pairs of observations, $(\boldsymbol{\theta}_1, \boldsymbol{\phi}_1)$, $(\boldsymbol{\theta}_2, \phi_2)$, …, $(\boldsymbol{\theta}_N, \boldsymbol{\phi}_N)$. For determining if an association is significant, we hypothesize H_o: No correlation or association exists between $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$. If the hypothesis is true, any pairing of the observed $\boldsymbol{\Theta}$ and $\boldsymbol{\Phi}$ values would give a calculated correlation near zero. Hence, if we randomly shuffled the observations $\boldsymbol{\theta}_i$, but kept the same order of values of $\boldsymbol{\phi}_i$, we would obtain a second calculated value of association for the sample. If H_o is true, that measure would be expected near zero. Repeating this for all possible orderings of $\boldsymbol{\theta}_i$ would give all possible correlation values for the observations, thereby simulating the distribution of the correlation measure *assuming* H_o *is true*.

The definition of a Type 1 error in significance testing is that we reject a true hypothesis H_o, with $\boldsymbol{\alpha}$ being the probability of making such an error. Hence, extreme values in the calculated distribution would represent random occurrences that would lead to a Type 1 error. We pick the $100\boldsymbol{\alpha}\%$ of the most extreme values to define a cutoff for the test statistic or coefficient; reject H_o if the original sample gave a result in this extreme (critical) region. That is, an overly large (or small) value of the coefficient from the original sample (compared to the extreme cutoff value) is used to reject the hypothesis of non-association.

There are $N!$ possible ways to pair the observed $(\boldsymbol{\theta}_i, \boldsymbol{\phi}_i)$ values. Ideally, we would evaluate all possible permutations of the observations, but this would imply millions of calculations, even for small N. Instead, we randomly select a smaller number of permutations of the data and use these to estimate the distribution. The script allows the user to specify the number, *NTrials*, of these trials. *NTrials* can be as small as 100 (primarily for testing) or as large as 10,000; the script defaults to 2500. Resampling is done by **`Vector_Corr`** when $N \leqslant 30-35$.

4. Other information

4.1. Material in IAMG server

The IAMG server contains the following:

- The script **`Vector_Corr`**, consisting of 30 MATLAB m-files (scripts and functions). These m-files may be modified by users for their own needs.
- A file titled **READ_ME_CORR.txt** that describes operation of the script. It also describes the format of the input data file.
- A file titled **WindOzoneData.txt** that contains 19 observations of ozone concentration and wind direction at a weather station in Milwaukee. This data is in Fisher (1993, Appendix B.18); see also Johnson and Wehrly (1977, Table 1). The data array is shown in Table 1.
- A file titled **WindOzoneResults.txt** that contains the text file generated by analysis of **WindOzoneData.txt**.
- A file titled **FluvialReachXbedData.txt** that contains 12 pairs of measurements of (1) the orientations of approximately straight reaches in the "latest" stages of the paleo-river that deposited the Rocktown channel sandstone (Cretaceous Dakota Formation, central Kansas), and (2) the vector means of cross bed orientations measured in the reaches. The variables are in columns 3 and 4, respectively, of Table 2. Fig. 2 shows that the two variables are well correlated. Data are from Siemers (1976, Table 4).
- A file titled **FluvialReachXbedResults.txt** that contains the text file generated by analysis of **FluvialReachXbedData.txt**.
- A file titled **PairedWindData.txt** that contains 21 pairs of measurements of wind directions taken at 6 a.m. and 12 noon at a weather station in Milwaukee (Johnson and Wehrly, 1977, Table 2; also in Fisher, 1993, Appendix B.21). Fig. 3 shows that the two variables are somewhat correlated, as seen by the 45° trends in the plot.
- A file titled **PairedWindResults.txt** that contains the text file generated by analysis of **PairedWindData.txt**.
- A file titled **UplandGravelsCorrelData.txt** that contains 63 sets of observations on gravels located in southern Maryland, probably deposited by the ancestral, south-to-southeast flowing Potomac River. The variables in the data set include (column 1) modal size of gravel fraction (phi units), (2) quartzite content in 16–32 mm fraction (percent), (3) chert content in 16–32 mm fraction (percent), and (4) azimuthal direction of paleocurrent indicators (Schlee, 1957, Figs. 12, 13, 17, 18). Schlee mapped his observations onto a grid by use of moving averages, from which the values in the data set were taken.
- A file titled **UplandGravelsCorrelResult.txt** that contains the text file generated by analysis of columns 1, 2, and 4 of **UplandGravelsCorrelData.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 bidirectional lines; these phenomena arise from processes that leave evidence of orientation but do not distinguish between directions $\boldsymbol{\theta}$ and $\boldsymbol{\theta}+180^\circ$. 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.

Krumbein (1939) and Fisher (1993, p. 37) recommend the following process for analyzing axial data. First, convert the axial directions $\boldsymbol{\psi}$ to vectors $\boldsymbol{\theta}$ by $\boldsymbol{\theta}_i = 2\boldsymbol{\psi}_i$ [mod 360] for all *i*. Second, analyze the resulting vectors using the methods described in this paper. No further transforms of the results are needed.

References

Blaesild, P., Granfeldt, G., 2003. Statistics with Applications in Biology and Geology. Chapman & Hall London, CRC, Boca Raton, Florida 555pp.

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

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.

Johnson, R.A., Wehrly, T.E., 1977. Measures and models for angular correlation and angular-linear regression. Journal Royal Statistical Society, Series B 39 (2), 222–229.

Jones, T.A., 2005. MATLAB functions to analyze directional (azimuthal) data—I: Single-sample inference. Computers & Geosciences, in press, doi:10.1016/j.cageo.2005.06.009.

Jupp, P.E., Mardia, K.V., 1980. A general correlation coefficient for directional data and related regression problems. Biometrika 67 (1), 163–173 Correction, 1981, 68 (3), 738.

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

Liddell, I.G., Ord, J.K., 1978. Linear–circular correlation coefficients: some further results. Biometrika 65 (2), 448–450.

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

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

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

Schlee, J., 1957. Upland gravels of southern Maryland. Bulletin Geological Society of America 68 (10), 1371–1410.

Siemers, C.T., 1976. Sedimentology of the Rocktown channel sandstone, upper part of the Dakota Formation (Cretaceous), central Kansas. Journal of Sedimentary Petrology 46 (1), 97–123.