Multivariate Analysis IV

Paruchuri R. Krishnaiah
Editor

PROBLEMS OF ASSOCIATION FOR BIVARIATE CIRCULAR DATA AND A NEW TEST OF INDEPENDENCE

Madan L. PURI* and J.S. RAO**

Indiana University, Bloomington, Ind., U.S.A.

1. Introduction and summary

Statistical data where the observations are directions in either two or three dimensions occur naturally in many diverse fields such as geology, biology, astronomy and medicine among others. These directions may be represented as unit vectors, that is, as points on the circumference of the unit circle if two dimensional or on the unit sphere in case they are three dimensional. These directions may also be represented in terms of angles with respect to a fixed "zero direction". It is natural to require that statistical techniques for such directional data have to be independent of this arbitrary chosen zero direction, as well as the sense of rotation, that is, whether one takes clockwise or anticlockwise as the positive direction. These natural restrictions rule out the application of most of the standard statistical methods for directional data. For instance, even the usual "Arithmetic Mean" and the "Standard Deviation" fail to be meaningful measures of location and dispersion. This novel area of statistics has been receiving increasing attention only recently and most of the statistical developments in this area have occurred during the past two decades, especially after the appearance of a paper by Fisher (1953). For a general survey of this field, the reader is referred to Mardia (1972) and Batschelet (1965).

In this paper we shall restrict our attention to circular data, that is, data on two dimensional directions. In many instances, we may measure more than one direction corresponding to each "individual" in the population either because we wish to gain more information than what a single

measurement could give or because we wish to study the interrelationships, if any (for example, correlation and regression), between such variates. Also, one might be interested in regression problems like predicting the paleocurrent direction on the basis of cross-bedding dip directions or pebble orientations, or in problems of correlation like measuring the association between the flight directions of pigeons and the prevalent wind direction. This important area of multivariate directional data analysis has not so far received much attention, with the result that the research work (on multivariate situations) has been very limited. To acquaint the reader with some background material, we give in section two, a brief review of some of the literature on association and independence for bivariate circular data. In section three, we introduce a new test of independence for measurements on a torus, that is, bivariate circular data. The asymptotic distribution theory of the proposed test statistic is derived in the last section.

2. Problems of association and independence

As stated in section one, we give here a brief survey of the papers of Downs (1974), Mardia (1975), Gould (1969) and Rothman (1971), all of which deal with problems of association for angular variables. While the first two of these papers attempt to define a measure of correlation for angular variates, the third discusses problems of regression, and the last introduces a test of independence. In what follows, $(X_1, Y_1), \ldots, (X_n, Y_n)$ will denote a random sample of observations on a torus, that is, X_i as well as Y_i are angles in $[0, 2\pi)$.

Downs (1974) defines a measure of "rotational correlation" for circular data which is in many ways analogous to the product moment correlation on the line. Let

$$\bar{C}_x = n^{-1} \sum_{i=1}^{n} \cos X_i, \qquad \bar{S}_x = n^{-1} \sum_{i=1}^{n} \sin X_i,$$

$$\bar{C}_y = n^{-1} \sum_{i=1}^{n} \cos Y_i, \qquad \bar{S}_y = n^{-1} \sum_{i=1}^{n} \sin Y_i,$$

be the arithmetic means of cosine and sine components of X and Y. Let

$$T = \frac{1}{n} \sum_{i=1}^{n} \begin{pmatrix} \cos Y_i - \bar{C}_y \\ \sin Y_i - \bar{S}_y \end{pmatrix} \begin{pmatrix} \cos X_i - \bar{C}_x \\ \sin X_i - \bar{S}_x \end{pmatrix}'.$$

Set

$$S_x^2 = 1 - (\bar{C}_x^2 + \bar{S}_x^2), \qquad S_y^2 = 1 - (\bar{C}_y^2 + \bar{S}_y^2),$$

and

$$S_{xy} = |T(T'T)^{-\frac{1}{2}}| \cdot \text{tr}(T'T)^{\frac{1}{2}}$$

where $(T'T)^{\frac{1}{2}}$ is the unique symmetric positive definite matrix whose square is $(T'T)$ and $(T'T)^{-\frac{1}{2}}$ is its inverse. Note that the determinant $|T(T'T)^{-\frac{1}{2}}|$ is ± 1 because of the orthogonality of the matrix. Also if R_x^2 and R_y^2 denote the squared lengths of the resultants corresponding to the directions $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$ respectively, then it is easily seen that

$$S_x^2 = 1 - R_{x/n^2}^2 \quad \text{and} \quad S_y^2 = 1 - R_{y/n^2}^2$$

which are the commonly used measures of variation. A justification for the rather complex definition of S_{xy} is not hard to find and the reader is referred to Downs (1974). Downs defines the circular rotational correlation $\gamma_c = S_{xy}/S_x \cdot S_y$. It can be seen that γ_c lies between -1 and $+1$ attaining one of the extreme values only when the Y-deviation (from its resultant direction) is a constant multiple of an orthogonal transformation (that is, a rotation) of the X-deviation for every pair (X_i, Y_i) in the sample. One has to keep in mind however that this may not be an appropriate measure of association when the correlation is not strictly rotational. γ_c remains invariant if the origins are changed for the X and Y measurements. The sampling distribution of γ_c has not been investigated so far and this limits the use of γ_c for statistical inference.

Mardia's (1975) correlation coefficient for circular data is based on the ranks of the observations, and is defined as follows: In analogy with the linear situation, a perfect correlation is said to exist between X and Y if the whole probability mass is concentrated on

$$(lX \pm mY + Z) \bmod 2\pi = 0$$

for some positive integers l, m and a fixed angular quantity Z. Define $X_i^* = lX_i \pmod{2\pi}$, $Y_i^* = mY_i \pmod{2\pi}$, and let the linear ranks of $X_1^*, \ldots, X_n^*$ be $1, \ldots, n$ respectively and those of $Y_1^*, \ldots, Y_n^*$ be $r_1, \ldots, r_n$ respectively. The angles (X_i^*, Y_i^*) are then replaced by the uniform scores $(2\pi i/n, 2\pi r_i/n)$. Now let R_1^2 and R_2^2 denote the lengths of the resultants corresponding to the directions $2\pi(i - r_i)/n$, $i = 1, \ldots, n$ and $2\pi(i + r_i)/n$, $i = 1, \ldots, n$ respectively. Mardia (1975) defines

$$\gamma_0 = \max(R_1^2/n^2, \ R_2^2/n^2)$$

as the circular rank correlation coefficient. It is clear that γ_0 lies between

zero and one, and that it remains invariant under changes of zero-directions of X and Y. We have $R_1^2/n^2 = 1$ for perfect "positive" dependence, and $R_2^2/n^2 = 1$ for perfect "negative" dependence. γ_0 will be close to zero if X and Y are uncorrelated. Mardia (1975) also discusses the asymptotic null distribution of γ_0 and gives a table of critical values.

Gould (1969), on the other hand, considers an analogue of the normal theory linear regression for circular variables. Let (x_i, Y_i), $i = 1, \ldots, n$ be observations on directions such that $Y_i, i = 1, \ldots, n$ are independently distributed as circular normal random variables $CN(\alpha + \beta x_i, \mathcal{K})$, that is, with density

$$f(y_i) = \frac{1}{2\pi I_0(\mathcal{K})} \exp(\mathcal{K} \cos(y_i - \alpha - \beta x_i))$$

where $x_1, \ldots, x_n$ are known numbers while α, β and $\mathcal{K}$ are unknown parameters. Since the logarithm of the likelihood function is

$$- n \log 2\pi - n \log I_0(\kappa) + \mathcal{K} \sum_{i=1}^{n} \cos(Y_i - \alpha - \beta x_i),$$

the MLE's (Maximum Likelihood Estimates) $\hat\alpha$ and $\hat\beta$ of α and β are the solutions of the equations

$$\sum_{i=1}^{n} \sin(Y_i - \hat\alpha - \beta x_i) = 0,$$

$$\sum_{i=1}^{n} x_i \sin(Y_i - \hat\alpha - \hat\beta x_i) = 0.$$

These solutions can be obtained by an iterative procedure discussed in Gould (1969). After solving for $\hat\alpha$ and $\hat\beta$, the MLE $\hat{\mathcal{K}}$ of $\mathcal{K}$ is obtained from the equation

$$\frac{I_1(\hat{\mathcal{K}})}{I_0(\hat{\mathcal{K}})} = \sum_{i=1}^{n} \cos(Y_i - \hat\alpha - \hat\beta x_i)/n.$$

Finally, for testing independence, Rothman (1971) assumes that the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ are from a continuous d.f. (distribution function) $F(x, y)$ whose marginal d.f.'s are $F_1(x)$ and $F_2(y)$. The problem is to test the hypothesis $H_0 : F(x, y) = F_1(x) \cdot F_2(y) \forall (x, y)$. With respect to the given origins for the two variates, let

$$F_{n1}(x) = n^{-1} \sum_{i=1}^{n} I(X_i, x),$$

iant under changes of zero-
for perfect "positive" depen-
dependence. γ_0 will be close to
dia (1975) also discusses the
s a table of critical values.
rs an analogue of the normal
es. Let (x_i, Y_i), $i = 1, \ldots, n$ be
$= 1, \ldots, n$ are independently
ables $CN(\alpha + \beta x_i, \mathcal{K})$, that is,

$- \beta x_i))$

le α, β and $\mathcal{K}$ are unknown
elihood function is

$(Y_i - \alpha - \beta x_i)$,

s) $\hat{\alpha}$ and $\hat{\beta}$ of α and β are the

erative procedure discussed in
MLE $\hat{\mathcal{K}}$ of $\mathcal{K}$ is obtained from

man (1971) assumes that the
a continuous d.f. (distribution
$_1(x)$ and $F_2(y)$. The problem is
$F_2(y) \forall (x, y)$. With respect to

$$F_{n2}(y) = n^{-1} \sum_{i=1}^{n} I(Y_i; y),$$

$$F_n(x, y) = n^{-1} \sum_{i=1}^{n} I(X_i; x) I(Y_i; y),$$

where

$$I(s; t) = \begin{cases} 1 & \text{if} \quad s \leq t, \\ 0 & \text{if} \quad s > t. \end{cases}$$

Thus $F_{n1}(x)$, and $F_{n2}(y)$ and $F_n(x, y)$ are the empirical d.f.'s of the X's, Y's and (X, Y)'s respectively. Distribution free tests of the form

$$\int \int T_n^2(x, y) \, dF_n(x, y), \quad \text{where} \quad T_n(x, y) = [F_n(x, y) - F_{n1}(x) F_{n2}(y)]$$

were considered earlier by Blum, Kiefer and Rosenblatt (1961) and since they are not invariant under different choices of the origins for X and Y, they are not applicable to the circular case. To circumvent this problem, Rothman (1971) suggested the modified statistic

$$C_n = n \int_0^{2\pi} \int_0^{2\pi} Z_n^2(x, y) \, dF_n(x, y),$$

where

$$Z_n(x, y) = \left[T_n(x, y) + \int \int T_n(x, y) \, dF_n(x, y) \right.$$
$$\left. - \int T_n(x, y) \, dF_{n1}(x) - \int T_n(x, y) \, dF_{n2}(y) \right].$$

The statistic C_n has the desired invariance property and its asymptotic distribution theory under the null hypothesis has been derived by Rothman (1971).

3. A new test for co-ordinate independence of circular data

Let $\mathcal{F}$ denote the family of probability distributions on the circumference $[0, 2\pi)$ of the circle with the property $F(\alpha + 2\pi) = F(\alpha) + \frac{1}{2}$ for all α. For instance circular distributions with axial symmetry would be in this class. In this section we propose a test which is applicable to testing independence when the marginals F_1 and F_2 belong to this class $\mathcal{F}$. When dealing with axial data, each observed axis will be represented by both its antipodal points for the purposes of this test. The proposed test may also be applied for testing independence in the non-axial case. But in this case, corresponding to every observed direction, we should add its antipodal

point also to the data thus doubling the original sample size. The effect such "doubling" would have on the power of the test procedure, is presently being investigated. Thus from now on, the random sample $(X_1, Y_1), \ldots, (X_n, Y_n)$ referred to in Sections 3 and 4 corresponds to the axial data with both ends represented as two distinct sample points or the "doubled" sample in the general non-axial case. This ensures that the marginal distributions F_1 and F_2 belong to the class $\mathscr{F}$.

As before, let $(X_1, Y_1), \ldots, (X_n, Y_n)$ denote a random sample of angular variates on the basis of which we wish to test the null hypothesis of independence. For any fixed x in $[0, 2\pi)$, let $N_1(x)$ denote the number of X_i's that fall in the half-circle $[x, x + \pi)$. Similarly, for any fixed y in $[0, 2\pi)$, let $N_2(y)$ denote the number of Y_i's which fall in the half circle $[y, y + \pi)$. Also, let $N(x, y)$ denote the number of observations (X_i, Y_i) that fall in the quadrant $[x, x + \pi) \times [y, y + \pi)$. Now defining the indicator variables

$$I_i(x) = \begin{cases} 1 & \text{if} \quad X_i \in [x, x + \pi), \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

and

$$\bar{I}_i(y) = \begin{cases} 1 & \text{if} \quad Y_i \in [y, y + \pi], \\ 0 & \text{otherwise} \end{cases} \tag{3.2}$$

we obtain

$$N_1(x) = \sum_{i=1}^{n} I_i(x), \qquad N_2(y) = \sum_{i=1}^{n} \bar{I}_i(y),$$

$$N(x, y) = \sum_{i=1}^{n} I_i(x) \bar{I}_i(y). \tag{3.3}$$

If the hypothesis of independence holds, then we should have

$$N(x, y) = N_1(x) \cdot N_2(y)/n$$

by the usual arguments. Thus we define

$$D_n(x, y) = n^{-\frac{1}{2}}[N(x, y) - N_1(x) N_2(y)/n] \tag{3.4}$$

as a measure of discrepancy between the observed and expected (under the hypothesis of independence) frequencies. Since $T_n(x, y)$ depends specifically on the choices of x and y, we suggest the (invariant) statistic

$$T_n = \int_0^{2\pi} \int_0^{2\pi} D_n^2(x, y) \frac{\mathrm{d}x}{2\pi} \frac{\mathrm{d}y}{2\pi} \tag{3.5}$$

$$= \frac{1}{4\pi^2 n} \int_0^{2\pi} \int_0^{2\pi} \left[N(x, y) - \frac{N_1(x) N_2(y)}{n} \right]^2 \mathrm{d}x \, \mathrm{d}y$$

nal sample size. The effect
of the test procedure, is
y on, the random sample
and 4 corresponds to the
istinct sample points or the
ase. This ensures that the
e class $\mathcal{F}$.

random sample of angular
est the null hypothesis of
$N_1(x)$ denote the number of
ly, for any fixed y in $[0, 2\pi)$,
in the half circle $[y, y + \pi)$.
tions (X_i, Y_i) that fall in the
the indicator variables

(3.1)

(3.2)

$),$

(3.3)

en we should have

$n]$ (3.4)

ved and expected (under the
ce $T_n(x, y)$ depends specifi-
he (invariant) statistic

(3.5)

$\dfrac{N_2(y)}{}\Bigg]^2 \, \mathrm{d}x \, \mathrm{d}y$

for testing independence. The integrand $D_n^2(x, y)$ is much like the usual chi-square test for independence from a 2×2 table. We now derive a computional form for T_n in terms of the X and Y spacings.

In view of (3.1), (3.2) and (3.3), we have

$$
\begin{aligned}
n D_n^2(x, y) =& \left[\left(1 - \frac{1}{n}\right)\left(\sum_{i=1}^n I_i(x)\bar{I}_i(y)\right) - \frac{1}{n}\left(\sum_{i \neq j} I_i(x)\bar{I}_j(y)\right)\right]^2 \\
=& \left(1 - \frac{1}{n}\right)^2 \left\{\sum I_i \bar{I}_i + \sum I_{ij}\bar{I}_{ij}\right\} \\
& - \frac{2}{n}\left(1 - \frac{1}{n}\right)\left\{\sum I_i \bar{I}_{ij} + \sum I_{ij}\bar{I}_i + \sum I_{ij}\bar{I}_{ik}\right\} \\
& + \frac{1}{n^2}\left\{\sum I_i \bar{I}_j + \sum I_{ij}\bar{I}_k + \sum I_i \bar{I}_{jk} + \sum I_{ij}\bar{I}_{k1}\right\}
\end{aligned}
$$

where $I_{ij} = I_i(x)I_j(x)$, $\bar{I}_{ij} = \bar{I}_i(y)\bar{I}_j(y)$, and summations are over all distinct subscripts. It is easy to check that

$$
\int_0^{2\pi} I_i(x)\mathrm{d}x = \int_0^{2\pi} \bar{I}_i(y)\mathrm{d}y = \pi,
$$

$$
\int_0^{2\pi} I_i(x)I_j(x)\mathrm{d}x = \pi - D_{ij},
$$

$$
\int_0^{2\pi} \bar{I}_i(y)\bar{I}_j(y)\mathrm{d}y = \pi - \bar{D}_{ij},
$$

where D_{ij} and $\bar{D}_{ij}$ are the "circular" distances between (X_i, X_j) and (Y_i, Y_j) respectively. Omitting the routine computations, we obtain

$$
\begin{aligned}
T_n =& \frac{1}{4\pi^2 n}\Bigg[(n-1)\pi^2 - \frac{\pi}{n}\left\{\sum(\pi - D_{ij}) + \sum(\pi - \bar{D}_{ij})\right\} \\
& + \left(1 - \frac{1}{n}\right)^2\left\{\sum(\pi - D_{ij})(\pi - \bar{D}_{ik})\right\} \\
& - \frac{2}{n}\left(1 - \frac{1}{n}\right)\left\{\sum(\pi - D_{ij})(\pi - \bar{D}_{ik})\right\} \\
& + \frac{1}{n^2}\left\{\sum(\pi - D_{ij})(\pi - \bar{D}_{k1})\right\}\Bigg],
\end{aligned}
$$

again the summations run over all the distinct subscripts. The statistic T_n being a function of the circular distances $\{D_{ij}\}$ and $\{\bar{D}_{ij}\}$, is clearly invariant under rotations of either coordinate axis.

4. The asymptotic null distribution

To derive the asymptotic null distribution of T_n, we will utilize the methods of Fourier analysis similar to those in Rao (1972) and Rothman (1971). Since $D_n(x, y)$ is a doubly periodic function (in both x and y) we may find the Fourier expansion

$$D_n(x, y) = \sqrt{n} \sum_{k=-\infty}^{\infty} \sum_{m=-\infty}^{\infty} Z_{km}\, e^{-ikx}\, e^{-imy}, \tag{4.1}$$

where

$$Z_{km} = \frac{1}{4\pi^2} \int_0^{2\pi} \int_0^{2\pi} \left[\frac{N(x, y)}{n} - \frac{N_1(x)}{n} \cdot \frac{N_2(y)}{n} \right] e^{ikx}\, e^{imy}\, dx\, dy \tag{4.2}$$

$$= \frac{1}{n} \sum_{j=1}^{n} \left(\int_0^{2\pi} I_j(x) \frac{e^{ikx}}{2\pi}\, dx \right) \left(\int_0^{2\pi} \bar{I}_j(y) \frac{e^{imy}}{2\pi}\, dy \right)$$

$$- \frac{\sum_{j=1}^{n} \left(\int_0^{2\pi} I_j(x) \frac{e^{ikx}}{2\pi}\, dy \right)}{n} \times \frac{\sum_{j=1}^{n} \left(\int_0^{2\pi} \bar{I}_j(y) \frac{e^{imy}}{2\pi}\, dy \right)}{n}.$$

From the definitions of $I_j(x)$ and $\bar{I}_j(y)$, it follows that

$$Z_{km} = \begin{cases} 0, & \text{if } k \text{ or } m \text{ is even,} \\[2mm] \frac{1}{\pi^2 mk} \left[\left(\frac{\sum_{j=1}^{n} e^{ikX_j}}{n} \right) \left(\frac{\sum_{j=1}^{n} e^{imY_j}}{n} \right) - \left(\frac{\sum_{j=1}^{n} e^{ikX_j + imY_j}}{n} \right) \right], \\[4mm] & \text{if both } k \text{ and } m \text{ are odd.} \end{cases}$$

Thus from (3.5), (4.1) and an application of Parseval's theorem, we have

$$T_n = n \sum_k \sum_m |Z_{km}|^2.$$

If can be verified that under the null hypothesis of independence

$$\mathbb{E} Z_{km} = 0, \qquad \forall k, m$$

$$\mathbb{E} Z_{km} Z_{k'm'} = \begin{cases} \frac{\delta_{kk'}\delta_{mm'}}{\pi^4 k^2 m^2} \left(\frac{n-1}{n^2} \right), & \text{if both } k \text{ and } m \text{ are odd,} \\ 0, & \text{otherwise} \end{cases}$$

where δ_{jk} is the usual Kronecker delta. Thus, the random Fourier coefficients Z_{km} have zero means and are uncorrelated for distinct pairs

(k, m). It may be remarked here that the above expectations may be calculated under the assumption that the X_i's and Y_i's have a uniform distribution, which under the null hypothesis are further independent, in view of the fact we could use a probability integral transformation as in Blum, Kiefer and Rosenblatt (1961). This transformation does not change the statistic T_n since the numbers $N(x, y)$, $N_1(x)$, $N_2(y)$ remain unchanged under monotone transformations.

Now by arguments similar to those in Rao (1972) or Rothman (1971), we have the result that T_n has asymptotically the same distribution as the sum of the squares of independent normal variables X_{km} with zero means and variances $(\pi^4 k^2 m^2)^{-1}$ for odd k and m. Thus the asymptotic characteristic function of T_n is given by

$$\varphi(t) = \prod_{k \text{ odd}} \prod_{m \text{ odd}} \left(1 - \frac{2it}{\pi^4 k^2 m^2}\right)^{-1/2}$$

$$= \prod_{k=1}^{\infty} \prod_{m=1}^{\infty} \left(1 - \frac{2it}{\pi^4 (2k-1)^2 (2m-1)^2}\right)^{-2}.$$

This characteristic function can be formally inverted as in Rao (1972). On the other hand by a result of Zolotarev (1961), if $F(x)$ denotes the asymptotic d.f. of T_n, then the upper tail probabilities relating to T_n may be approximated as follows:

$$\lim_{x \to \infty} \frac{1 - F(x)}{P[\chi_4^2 > \pi^4 x]} = \prod_{k=1}^{\infty} \prod_{m=1}^{\infty} \left(1 - \frac{1}{(2k-1)^2 (2m-1)^2}\right)^{-2}$$

$$(k, m) \neq (1, 1)$$

where χ_4^2 denotes a random variable having the chi-square distribution with 4 degrees of freedom.

Acknowledgment

References

[1] Batschelet, E. (1965). Statistical Methods for the Analysis of Problems in Animal Orientation and Certain Biological Rhythms, Amer. Inst. Biol. Sciences, Washington.
[2] Blum, J.R., Kiefer, J. and Rosenblatt, M. (1961). Distribution free test of independence based on the sample distribution function, Ann. Math. Statist. 32, 485–492.

[3] Downs, T.D. (1974). Rotational Angular Correlations, in *Biorhythms and Human Reproduction*, (M. Ferin et al. Eds.), John Wiley, New York: 97–104.

[4] Fisher, R.A. (1953). Dispersion on a sphere, *Proc. Roy. Soc.* London A**217**, 295–305.

[5] Gould, A.L. (1969). A regression technique for angular variates, *Biometrics*, **25**, 683–700.

[6] Mardia, K.V. (1972). *Statistics of Directional Data*, Academic Press, New York.

[7] Mardia, K.V. (1975). Statistics of Directional Data, to appear in *Jour. Roy. Statist. Soc.*, Series B.

[8] Rao, J. S. (1972). Some variants of chi-square for testing uniformity on the circle, *Z. Wahrscheinlichkeitstheorie Verw. Geb.* **22**, 33–44.

[9] Rothman, E.D. (1971). Tests of coordinate independence for a bivariate sample on torus, *Ann. Math. Statist.*, **42**, 1962–1969.

[10] Zolotarev, V.M. (1961). Concerning a certain probability problem, *Theor. Probability Appl*, 201–204.