Methods and Applications of Statistics *in*

ENGINEERING, QUALITY CONTROL, and the PHYSICAL SCIENCES

N. Balakrishnan, Editor

WILEY

14 Directional Statistics, I

S. Rao Jammalamadaka

14.1 Introduction

This chapter explores the novel area of directional statistics. In many diverse scientific fields, the measurements are directions. For instance, a biologist may be measuring the direction of flight of a bird or the orientation of an animal, while a geologist may be interested in the direction of the earth's magnetic pole. Such directions may be in two-dimensions as in the first two examples or in three-dimensions like the last one. A set of such observations on directions is referred to as *directional data*.

Two-dimensional directions can be represented as angles measured with respect to some suitably chosen "zero direction," i.e., the starting point, and a "sense of rotation," i.e., whether clockwise or counterclockwise, is taken as the positive direction. Since a direction has no magnitude, these can be conveniently represented as points on the circumference of a unit circle centered at the origin or as unit vectors connecting the origin to these points. Because of this circular representation, observations on such two-dimensional directions are also called *circular data*. Similarly, directions in three dimensions may be represented by two angles (akin to the representation of points on the earth's surface by their longitude and latitude), as unit vectors in three-dimensions, or as points on the surface of a unit sphere. Because of this, directional data in three-dimensions are also referred to as *spherical data*.

Directional data have many unique and novel features both in terms of modeling and in their statistical treatment. For instance, the numerical representation of a two-dimensional direction as an angle or a unit vector is not necessarily unique since the angular value depends on the choice of what is labeled as the zero-direction and the sense of rotation. What is considered 60^o by a mathematician who takes true East as the zero-direction and counterclockwise as the positive direction comes out to be 30^o to a geologist who takes true North as the zero and clockwise as the positive direction.

The same is true of the specific values assigned to any spherical data set. It is therefore important to make sure that our conclusions (i.e., data summaries, inferences, etc.) are a function of the given observations and not dependent on the arbitrary values by which we refer to them. That is, we should aim at conclusions that do not depend on the arbitrary choice of origin and sense of rotation. Again, because of this arbitrary choice, there is also no natural ordering or ranking of the observations since whether one direction is "larger" than

the other depends on whether clockwise or counterclockwise is treated as being the positive direction as well as where the "zero" is. This makes rank-based methods essentially inapplicable. Finally, since the "beginning" coincides with the "end" i.e., the data is periodic, methods for dealing with directional data should take careful note of how to measure the distance between any two points.

Such distinctive features make directional analysis substantially different from the standard "linear" statistical analysis of univariate or multivariate data that one finds in most statistics books. The need for "invariance" of statistical methods and measures with respect to the choice of this arbitrary zero-direction and sense of rotation makes many of the usual linear techniques and measures often misleading, if not entirely meaningless. Commonly used summary measures on the real line, such as the sample mean and variance, turn out to be inappropriate as do all the moments and cumulants. Analytical tools such as the moment generating function and other generating functions are equally useless. Many notions such as correlation and regression as well as their statistical measures need to be reinvented for directional data. Similarly, such ideas of statistical inference as unbiasedness, loss functions, variance bounds, etc., need to be redefined with caution.

All in all, this area of directional data provides an inquisitive reader with many open research problems and is a fertile area for developing new statistical methods and inferential tools. There is also an opportunity to develop new and novel "applications" to problems arising in natural, physical, medical, as well as social sciences.

It may be remarked that in studying circadian or other rhythms, the circle may be used to represent one cycle, and the interest may lie in the *timing* of an event within this cycle, say for instance when the body temperature or the blood pressure peaks within the day. Thus certain aspects in the study of biological rhythms provide an important example from biology that can be put into this circular data framework. Because biological rhythms control characteristics such as sleep-wake cycles, hormonal pulsatility, body temperature, mental alertness, reproductive cycles, and so on, there has been a renewed interest among medical professionals in such topics as chronobiology, chronotherapy, and the study of the biological clock.

The aim here is to make the readers, and especially applied scientists, aware of the limitations of the standard "linear" statistical methods so that they can correctly model, analyze, and make inferences on problems that have to do with directions. Since solutions to many directional data problems are nontrivial and often not obtainable in simple closed analytical form, related computer software is essential for practitioners to be able to use these methods. One such software called `CircStats`, based on `R`, is freely available.

Significant books on circular data include those by Batschelet [1], Fisher [4], Mardia and Jupp [11], and Jammalamadaka and SenGupta [7]. Readers interested in directions in three dimensions should consult the books by Watson [14] and Fisher et al. [3].

14.2 Representation and a Useful Model

A direction in two dimensions can be represented in any of the following equivalent ways: (i) as an angle $0 \leq \theta < 2\pi$, (ii) as a point on the circumference of a unit circle, (iii) as a vector of unit length $\mathbf{x} = (x_1 = \cos\theta,\ x_2 = \sin\theta)$, or (iv) as a complex number $z = e^{i\theta}$ of unit modulus. Similarly a three dimensional direction can also be represented in several equivalent ways, for example: (i) in terms of

angles, $(-\pi \le \phi < \pi,\ -\pi/2 \le \theta < \pi/2)$, called the *longitude (W-E)* and *latitude (S-N)* respectively, and (ii) as a unit vector—with rectangular coordinates,

$$\mathbf{x} = (x_1 = \cos\theta,\ x_2 = \sin\theta.\cos\phi,\ x_3 = \sin\theta.\sin\phi).$$

The unit vector representation readily generalizes to directions in any dimension.

One of the most useful and often-used models for directional data is due to Langevin [10], rediscovered by von Mises [13], and developed for three dimensions by Fisher [2]. The density toward any p-dimensional direction, represented as a vector of unit length $\mathbf{x}$ is given by

$$C_p(\kappa)e^{\kappa\mathbf{x}'\mu}, \qquad ||\mathbf{x}|| = 1, \quad ||\mu|| = 1,$$

where μ of unit length denotes the mean direction. It can be checked out that the normalizing constant $C_p(\kappa)$ is given by

$$C_p(\kappa) = \frac{\kappa^{\frac{p}{2}-1}}{(2\pi)^{\frac{p}{2}} I_{(\frac{p}{2}-1)}(\kappa)}.$$

In particular for the circular case ($p = 2$), this constant reduces to

$$C_2(\kappa) = \frac{1}{2\pi I_0(\kappa)},$$

where $I_0(\kappa)$ is the modified Bessel function of the first kind and order zero) giving the von Mises or the "Circular Normal" distribution, and for the spherical case ($p = 3$), this constant reduces to

$$C_3(\kappa) = \frac{\kappa}{4\pi \sinh(\kappa)},$$

yielding the distribution whose sampling properties were discussed by Fisher [2] in a fundamental paper. This distribution has many elegant properties, and the sampling distribution of the resultant vector

$$\sum \mathbf{x_i},$$

which happens to be a complete sufficient statistic for this family, can be obtained so that inference (estimation and hypothesis testing for parameters such as μ and κ) can be relatively easily handled.

Clearly there are large classes of other models. But their use for data analysis can be judged only by the availability of relevant sampling theory for inference. One large class of circular distributions, the so-called "wrapped" models, are obtained by wrapping any linear distribution around the circle. See, e.g., Gatto and Jammalamadaka [5], who discuss inference for a wrapped stable family of models, and Jones et al. [8], who discuss wrapping a t-distribution.

14.3 Nonparametric Methods

One- and multi-sample problems of inference that do not depend on specific models (i.e., model-free or nonparametric methods) are available for directional statistics. The classical one-sample problem of *goodness-of-fit* can be converted to one of testing uniformity just as one does on the real line, as long as one makes sure that the proposed test statistic is invariant under choice of origin and rotation. Such invariant versions of classical tests like the χ^2, the Kolmogorov–Smirnov, and the Cramer–von Mises tests have been developed by Rao [12], Kuiper [9], and Watson [14] respectively. Rao [12] considers the average of χ^2 over all possible choices of the zero direction, to obtain an invariant version of it. If $x_{(1)} \le \cdots \le x_{(n)}$ denote the ordered observations corresponding to $x_1, \ldots, x_n$, the empirical distribution function (edf) is defined by

$$F_n(x) = \begin{cases} 0 & \text{if } x < x_{(1)}, \\ i/n & \text{if } x_{(i)} \le x < x_{(i+1)}, \\ 1 & \text{if } x \ge x_{(n)}. \end{cases}$$

One can define an edf for the circular case analogously with respect to any arbitrary origin, but the values taken by such an

edf will depend on the choice of this origin as well as whether clockwise or counterclockwise is taken as the positive direction. Thus in order to be able to use these for circular data problems, statistics such as the Kolmogorov–Smirnov and Cramer–von Mises should be modified so as to make them rotation invariant.

Recall the one-sided Kolmogorov–Smirnov statistics given by

$$\begin{aligned} D_n^+ &= \sqrt{n}\sup_x(F_n(x) - F(x)), \\ D_n^- &= \sqrt{n}\sup_x(F(x) - F_n(x)) \\ &= -\sqrt{n}\inf_x(F_n(x) - F(x)). \end{aligned}$$

The more common two-sided Kolmogorov–Smirnov statistic D_n can then be written as

$$\begin{aligned} D_n &= \sqrt{n}\sup_n |F_n(x) - F(x)| \\ &= \max(D_n^+, D_n^-). \end{aligned}$$

Noticing that D_n^+ gains (or loses) just as much as D_n^- loses (or gains) due to a rotation, Kuiper [9] suggested the statistic

$$V_n = \left(D_n^+ + D_n^-\right).$$

Another classical alternative edf-based test for goodness-of-fit on the real line is provided by the Cramer–von Mises test, given by

$$C_n^2 = n\int_{-\infty}^{\infty}(F_n - F)^2\, dF.$$

Watson [14] provided an invariant modification of this that is suitable for circular data, and it is defined by

$$\begin{aligned} W_n^2 &= \int_{-\infty}^{\infty}\left[(F_n - F) - \right. \\ &\qquad \left.\int_{-\infty}^{\infty}(F_n - F)\, dF\right]^2 dF. \end{aligned}$$

Note that if the Cramer–von Mises statistic can be thought of as the "second moment" of $(F_n - F)$, Watson's statistic is similar to the expression for "variance" so that if the quantity $(F_n - F)$ changes by a constant δ due to a change of origin, as we stated before, the variance will not change.

Two circular samples can be compared by using two-sample versions of the Kuiper and Watson tests, among others, or by using the so-called "spacings-frequencies," which are the counts of the first sample that fall in between the gaps made by the second sample. See Holst and Rao [6] for a thorough discussion of this class of two-sample nonparametric tests.

14.4 Multivariate Problems

One can consider two or more directional variables simultaneously, or directional variables in conjunction with linear variables. For instance, models for bivariate circular data have the torus (circle × circle) as their sample space, whereas one circular and one linear variable take their values on the surface of a cylinder (circle x R^1). Questions of modeling, correlation, and regression in such a context are discussed in Fisher [4, Chap. 6], Mardia and Jupp [11, Chap. 11], and Jammalamadaka and SenGupta [7, Chap. 8].

14.5 Final Remarks

The aim of this chapter is somewhat limited and is mainly to bring awareness about the fact that the standard linear models and methods of analyses in which one is usually trained are not applicable when dealing with directional data. There are many excellent books on the topic of directional statistics to which we refer the interested reader for further details.

References

1. Batschelet, M. (1981). *Circular Statistics in Biology.* Academic Press, London.
2. Fisher, R. A. (1953). Dispersion on a sphere. *Proc. Roy. Soc., London, Ser. A.*, **217**, 295–305.
3. Fisher, N. I., Lewis, T., and Embleton, B. J. J. (1987). *Statistical Analysis of Spherical Data.* Cambridge University Press, Cambridge.
4. Fisher, N. I. (1993). *Statistical Analysis of Circular Data.* Cambridge University Press, Cambridge.
5. Gatto, R. and Jammalamadaka, S. Rao (2003). Inference for wrapped symmetric alpha-stable circular models. *Sankhya*, **65**, 333–355.
6. Holst, L. and Rao, J. S. (1980). Asymptotic theory for families of two-sample nonparametric statistics. *Sankhya, Ser. A*, **42**, 19–52.
7. Jammalamadaka, S. Rao, and SenGupta, S. (2001). *Topics in Circular Statistics.* World Scientific, Singapore.
8. Jones, M. C., Lewis, T., and Pewsey, A. (2007). The wrapped t-family of circular distributions. *Australian and New Zealand Journal of Statistics*, **49**, 79–91.
9. Kuiper, N. H. (1960), Tests concerning random points on a circle. *Ned. Akad. Wet. Proc.*, **63**, 38–47.
10. Langevin, P. (1905). Magnetisme et theorie des electrons. *Ann. Chim. Phys.*, **5**, 71-127.
11. Mardia, K. V. and Jupp, P. E. (2000). *Directional Statistics.* Wiley, New York.
12. Rao, J. S. (1972). Some variants of chi-square for testing uniformity on the circle. *Zeitschrift fur Wahrscheinlichkeitstheorie and vervandt Gebiete*, **22**, 33–44.
13. von Mises, R. (1918). Über die "Ganzzahligkeit" der Atomgewichhte und Verwandte Fragen. *Physikal. Z.*, **19**, 490–500.
14. Watson, G. S. (1961). Goodness-of-fit tests on the circle. *Biometrika.* **48**, 109–114.
15. Watson, G. S. (1983). *Statistics on Spheres.* Wiley, New York.