Algorithms for the Cyclic Mean and Median[1]

A.J. Allinger
Design Service Corporation
2013-2025

Abstract: This article will show how to represent cyclic numbers, and add, subtract, and average them. A new algorithm with complexity n log(n) is introduced for finding the median of a set of cyclic data . The proposed average will preserve not only phase information, but also lap information of the input. Source code in FORTRAN and C is provided.

Introduction

Circular numbers appear in many applications. Time of day and compass heading the two most important, but there are others: musical intervals, gear trains, and waveforms can be measured in terms of cyclic variables. What sets these measurements apart is that they *wrap* at some point and an increase past 12:59:59, for example, resets to 1:00:00. This makes operations as familiar as subtraction and averaging treacherous on cyclic quantities.

The average of cyclic quantities, such as time of day, cannot be computed by the means used for linear data. The usual approach is to treat the data as vectors on the unit circle, and return the arctangent of the resultant. This technique is appropriate for directional data in two dimensions.

Background

Musical notes are cyclic with period 12 (in the usual Western 12-tone equally tempered system). Low C sounds like high C. One systematic way to represent notes is by MIDI note number**.** In this system middle C is 60, and naturally B is 59. Their difference is 1. High C, 12 steps higher at

1 This article is adapted from “Circular Values Math and Statistics with FORTRAN” (2013), https://www.codeproject.com/Articles/695494/Circular-Values-Math-and-Statistics-with-FORTRAN

72, also sounds much like middle C. Treating note number as a cyclic quantity with period 12 makes every note equivalent to every note 12 steps apart. Thus, 72 ~ 60, and 60 - 59 = 1

But it's not that easy. The difference between 1 degrees and 359 degrees is 2, not 358. The real difference between the values is measured by the shortest path between them, and with cyclic variables, there are always two paths to consider. (Plus an infinite number of paths obtained by making full laps around the cycle!)

The basic operations will now be described:

Let t denote the period, and X the data. t shall be a positive number.

Phase

Various units are used to measure phase, such as degrees, radians, or cycles. Generally, phase may be measured in the same units as the input data. For example 1:00pm on a 24-hour period, has the phase of 13.0 hours. The phase p, of a measurement x, in period t is:

$p = x \% t$

To make comparisons possible, all phase data must lie in the same range. Let us choose the range 0...t. (-t/2 ... t/2 is another reasonable choice.) Negative data has a negative phase that should be brought into the standard range 0...t. That is:

$$p' = \begin{Bmatrix} p+t, & p<0 \\ p, & \text{otherwise} \end{Bmatrix}$$

Addition

Cyclic quantities add just like linear quantities.

$x = y + z$

Subtraction

Now things start to become more complicated. Two phase differences must be distinguished: a *directed difference* and an *absolute difference*. If direction is not important, the absolute difference can be found be subtraction, followed by a choice between the two paths around the circle. The absolute phase difference d_a between point A and point B is

$$d_m = |b-a| \% t$$
$$d_a = \min(d_m, t-d_m)$$

To preserve the direction, a little more care must be taken. The directed difference d_d from point A to point B is

$$d_m = (b-a)\%t$$

$$d_d = \begin{cases} d_m + t, & d_m < \frac{t}{2} \\ d_m - t, & d_m \geq \frac{t}{2} \\ d_m, & \text{otherwise} \end{cases}$$

The result lies in the interval $[-\frac{t}{2}, \frac{t}{2})$

Scalar Multiplication

Cyclic quantities may be multiplied by a scalar in the usual way:

x' = xs

If the purpose is to change the unit of measurement, then it is important to remember to scale the period also:

t' = ts

Cyclic Averages

Vector-sum Average

The following question was asked on a web forum at control.com:

> *Posted by Ville on 18 May, 2005 - 7:46 pm*
>
> *Hi! I have a weather station which measures both the direction and speed of the wind. It is connected to a SCADA system (TSX57xxx + Monitor Pro). Now I have a mathematical problem with calculating the average direction of the wind .....*

The most common method for finding the average of a cyclic quantity is the *vector-sum* method. http://en.wikipedia.org/wiki/Mean_of_circular_quantities The wind speeds s and directions h may be broken into vector components by:

$$x = \frac{1}{N}\sum s_i \cos(h_i)$$
$$y = \frac{1}{N}\sum s_i \sin(h_i)$$

Then the average speed and direction are given by the magnitude and angle

$$s_{avg} = \sqrt{x^2 + y^2}$$
$$h_{avg} = \arctan(y, x)$$

The vector sum of wind measurements taken at regular intervals yields the overall displacement of air in the neighborhood of the measuring station. Thus, in a physically meaningful sense, it can be said to be the correct average.

More generally, it doesn't have the properties that might be hoped for. For example, the average of {0, 0, 90 degrees} ought to be 30 degrees, but the vector sum method yields arctan(1, 2) = 26.56 degrees. Is there a way to represent the wrapping of cyclic variables without imposing an artificial 2-D structure on the problem?

The Cyclic Mean

Yes, in the same thread above, David Radcliffe gave a succinct solution:

> I propose the following algorithm for calculating the average of a set of directions, measured from 0 to 360 degrees:
>
> 1. Find the average and standard deviation of the given numbers.
> 2. Increase the smallest number by 360.
> 3. Repeat steps 1 and 2 until all numbers are greater than 360.
> 4. Choose the average that yields the smallest standard deviation.
> 5. If the average is greater than 360 then subtract 360.

If the circle is thought of as a line sewn together at a seam, Radcliffe's algorithm considers the intervals between all adjacent data points as a possible location for the seam. Since there are n "seams" one must consider, it is possible to have as many as an n-way tie for the least variance. This will occur if all the data points are equally spaced around the circle.

The algorithm was subsequently rediscovered by Olson (2011). He observed that by using the short formula for the variance: $\sigma^2 = \frac{\sum x^2}{N} - \left(\frac{\sum x}{N}\right)^2$

the sums and sums-of-squares need only to be updated at each step, with the formulas:

$$s_1' = s_1 + t$$
$$s_2' = s_2 + 2tx_i + t^2$$

The Cyclic Median

A simple extension of the foregoing results in a new, efficient algorithm for the cyclic median. Instead of minimizing the variance, the quantity to seek the minimum of is the mean absolute deviation: $\sigma_1 = \frac{1}{N}\sum |x_i - x_m|$

As with the linear median, the cases of odd and even *n* must be treated separately. Applying the definition of mean absolute deviation to the sorted data, for the odd case

$$n\sigma_1 = (x_m - x_1) + (x_m - x_2) + \cdots + (x_{n-1} - x_m) + (x_n - x_m)$$
$$n\sigma_1' = (x_m' - x_2) + (x_m' - x_3) + \cdots + (x_n - x_m') + (x_1 + t - x_m')$$

where if the median x_m, at one step is x_j, the median at the next step, $x_{m'}$ is x_{j+1} leading to the update formulas:

$$s_1' = s_1 + t + 2x_i - \begin{cases} x_{(n/2+i)\%n} + x_{(n/2+i+1)\%n}, & n\%2=1 \\ 2x_{(n/2+i)\%n}, & n\%2=0 \end{cases}$$

Substitute x_n for x_0 as required.

The limiting step is the n log(n) sorting step. Thus, finding the circular median is perfectly feasible even on large data sets.

Kanti Mardia's 1972 textbook defines the circular median in terms of diameter drawn across the circle that divides the data equally. His definition suggests an algorithm of comparable efficiency to the one described here. The new method has the advantage of conceptual similarity to the algorithm for the mean, and it also computes the mean absolute deviation.

Preserving cycle information

Cycle information is preserved. For example, taking the period to be 24 hours, averaging 30 and 32 hours yields 31, not 7. This is done by first computing the *linear* mean, then choosing the shortest distance to the cyclical (phase only) mean described above. This gives an answer correct both in phase and in magnitude--a potentially very valuable feature. Suppose that a drive pinion turns a much larger gear. Take one turn of the drive pinion as the period. Clearly, it is essential to keep track of the number of revolutions of the pinion to know the position of the gear.

Points of Interest

Sometimes a variable with a latent period can be made amenable to cyclic number treatment by rescaling the input. In music theory, the circle-of-fifths distance between two notes is how many steps of five semi-tones is required to move from one note to the other. Chord and key changes usually have short circle-of-fifths distances. Multiplying the MIDI note number by 5, and leaving the period at 12 makes the phase differences of the rescaled data equal the circle-of-fifths distance.

Implementation

Functions to compute the trigonometric average, the cyclic mean, the cyclic median, and the weighted cyclic mean have been implemented in FORTRAN with a complete C translation and are freely available at http://13olive.net/code/cyclic3.zip

References

Kanti V. Mardia, *Statistics of Directional Data* London: Academic Press, 1972. Cited in: Bennett Sango Otieno, *An Alternative Estimate of Preferred Direction for Circular Data*, Virginia Polytechnic Institute PhD dissertation, 2002.

David Radcliffe, 2005. https://control.com/forums/threads/calculating-an-average-value-of-the-wind-direction.20840/

"On computing the average orientation of vectors and lines," Edwin Olson, 2011. http://april.eecs.umich.edu/pdfs/olson2011orientation.pdf

"Circular Values Math and Statistics with C++11" , Lior Kogan, 2013. https://www.codeproject.com/Articles/190833/Circular-Values-Math-and-Statistics-with-Cplusplus

"True Circular Mean," David G. Long, 2023. https://mers.byu.edu/circularmean.html