Journal of Modern Applied Statistical Methods

Volume 4 | Issue 1 Article 9

5-1-2005

Effect Of Position Of An Outlier On The Influence Curve Of The Measures Of Preferred Direction For Circular Data

B. Sango Otieno
Grand Valley State University

Christine M. Anderson-Cook
Los Alamos National Laboratory

Journal of Modern Applied Statistical Methods
May, 2005, Vol. 4, No.1, 81-89

1538 – 9472/05/$95.00

Effect Of Position Of An Outlier On The Influence Curve Of The Measures Of Preferred Direction For Circular Data

B. Sango Otieno
Department of Statistics
Grand Valley State University

Christine M. Anderson-Cook
Statistical Sciences Group
Los Alamos National Laboratory

Circular or angular data occur in many fields of applied statistics. A common problem of interest in circular data is estimating a preferred direction and its corresponding distribution. It is complicated by the wrap-around effect on the circle, which exists because there is no natural minimum or maximum. The usual statistics employed for linear data are inappropriate for directional data, as they do not account for its circular nature. The robustness of the three common choices for summarizing the preferred direction (the sample circular mean, sample circular median and a circular analog of the Hodges-Lehmann estimator) are evaluated via their influence functions.

Key words: Circular distribution, directional data, influence function, outlier

Introduction

The notion of preferred direction in circular data is analogous to the center of a distribution for data on a linear scale. Unlike in linear data where a center always exists, if data are uniformly distributed around the circle, then there is no natural preferred direction. Therefore, it is appropriate and desirable that all sensible measures of preferred direction are undefined if the sample data are equally spaced around the circle. This article considers estimating the preferred direction for a sample of unimodal circular data. Three choices for summarizing the preferred direction are the mean direction, the median direction (Fisher 1993) and the Hodges-Lehmann estimate (Otieno & Anderson-Cook, 2003a).

B. Sango Otieno is Assistant Professor at Grand Valley State University. He is a member of Institute of Mathematical Statistics and Michigan Mathematics Teachers Association. Email: otienos@gvsu.edu. Christine Anderson-Cook is a Statistician at Los Alamos National Laboratory. She is a member of the American Statistical Association, and the American Society of Quality. Email: c-and-cook@lanl.gov.

The sample mean direction is a common choice for moderately large samples, because when combined with a measure of sample dispersion, it acts as a summary of the data suitable for comparison and amalgamation with other such information. The sample mean is obtained by treating the data as vectors of length one unit and using the direction of their resultant vector. Given a set of circular observations $\theta_1, \ldots, \theta_n$, each observations is measured as a unit vector with coordinates from the origin of $(\cos(\theta_i), \sin(\theta_i))$, i = 1, . . ., n. The resultant vector of these n unit vectors is obtained by summing them componentwise to get the resultant vector $R = \left(\sum_{i=1}^{n} \cos(\theta_i), \sum_{i=1}^{n} \sin(\theta_i) \right) = (C, S)$, say. The sample circular mean is the angle corresponding to the mean resultant vector $\overline{R} = \frac{R}{n} = \left(\frac{C}{n}, \frac{S}{n} \right) = (\overline{C}, \overline{S})$. That is, the angle corresponding to the mean resultant length $|\overline{R}| = \sqrt{(\overline{C}^2 + \overline{S}^2)}$.

Jamalamadaka and SenGupta (2001), show that the sample circular mean direction is location invariant, that is, if the data are shifted by a certain amount, the value of the sample

circular mean direction also changes by that amount.

An alternative, the sample median, can be thought of as the location of the circumference of the circle that balances the number of observations on the two halves of the circle, Otieno and Anderson-Cook (2003b). The sample median direction $\tilde{\theta}$ of angles $\theta_1, \ldots, \theta_n$, is defined to be the point P on the circumference of the circle that satisfies the following two properties: (a) The diameter PQ through P divides the circle into two semi-circles, each with an equal number of observed data points and, (b) the majority of the observed data are closer to P than to the anti-median Q. See Mardia (1972, p.28-30) and Fisher (1993, p. 35-36).

Note, the antimedian can be thought of as the meeting point of the two tails of the distribution on the opposite side of the circle. Intuitively, fewer observations are expected at the tails. As with the linear case, for odd size samples the median is an observation, while for even sized samples the median is the midpoint of two adjacent observations. Observations directly opposite each other do not contribute to the preferred direction, since in such a case the observations balance each other for all possible choices of medians. The procedure for finding the circular median has the flexibility to find a balancing point for situations involving ties, by mimicking the midranking idea for linear data.

Otieno and Anderson-Cook (2003b) describe a strategy for more efficiently dealing with non-unique circular median estimates especially for small samples, which are commonly encountered in circular data. Note that the angle $\tilde{\theta}$ which has the smallest circular mean deviation given by $d(\tilde{\theta}) = \pi - \frac{1}{n}\sum_{i=1}^{n}\left|\pi - \left|\theta_i - \tilde{\theta}\right|\right|$ is the circular median, Fisher (1993).

A third measure of preferred direction for circular data is the circular Hodges-Lehmann estimate of preferred direction, subsequently referred to as HL. This is the circular median of all pairwise circular means of the data (Otieno & Anderson-Cook, 2003a). As with the linear case, there are three possible methods for calculating this quantity based on which pairs of observations are considered.

The three possible methods involve using the circular means of all distinct pairs of observations, all distinct pairs of observations plus the individual observations (which are essentially pairwise circular means of individual observations with themselves), and all possible pairwise circular means. The estimates obtained by all the three methods, divide the obtained pairwise circular means evenly on the two semicircles. All these estimates of preferred direction are location invariant, since they satisfy the definition of the circular median, which is also location invariant. The approach used is feasible regardless of sample size or the presence of ties. Note that no ranking is used in computing the new measure, since on the circle there is no uniquely defined natural minimum/maximum. Simulation results show that the three HL measures tend towards being asymptotically identical, Otieno (2002), as is the case of linear data, Huber (1981).

Three choices are presented for estimating preferred direction for a single population of circular measures, and study their robustness via their influence curve. As with linear data, where the mean and the median represent different types of centers for data sets, the three estimates of preferred direction also have relative trade-offs for what they are trying to estimate as well as how they deal with lack of symmetry and outliers. The following data set is considered, which give a small overview of the types of data that may be encountered in practice by say, biologists. The data given in Table 1, relates the homing ability of the Northern cricket frog, *Acris crepitans*, as studied by Ferguson, et. al. (1967).

Table 1: Frog Data-Angles in degrees measured due North.

104	110	117	121	127	130	136	145	152
178	184	192	200	316				

Methodology

A circular distribution (CD) is a probability distribution whose total probability is concentrated on the circumference of a unit circle. A set of identically distributed independent random variables from such a distribution is referred to as a random sample from the CD. See Jammalamadaka & SenGupta (2001, p. 25-63) for a detailed discussion of circular probability distributions. Two frequently used families of distributions for circular data include the von Mises and the Uniform distribution.

The von Mises distribution VM (μ, κ), is a symmetric unimodal distribution characterized by a mean direction μ,and concentration parameter κ, with probability density function $f(\theta) = [2\pi I_0(\kappa)]^{-1} \exp[\kappa \cos(\theta - \mu)]$, $0 \le \theta, \mu < 2\pi$ and $0 \le \kappa < \infty$, where $I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp[\kappa xos(\phi)]d\phi = \sum_{j=0}^{\infty} -\frac{\kappa^{2j}}{4^j j^2}$ is the modified Bessel function of order zero.

The concentration parameter, κ, quantifies the dispersion. If κ is zero, $f(\theta) = \frac{1}{2\pi}$ and the distribution is uniform with no preferred direction. As κ increase from zero, $f(\theta)$ peaks higher about μ. The von Mises is symmetric since it has the property $f(\mu + \theta) = f(\mu - \theta)$, for all θ, where addition or subtraction is modulo 2π. With the uniform or isotropic distribution, however, the total probability is spread out uniformly on the circumference of a circle; that is, all directions are equally likely. It thus represents the state of no preferred direction.

The von Mises is similar in importance to the Normal distribution on the line, (Mardia,1972). When $\kappa \ge 2$, the von Mises distribution VM (μ, κ), can be approximated by the Wrapped Normal distribution WN(μ, ρ), which is a symmetric unimodal distribution obtained by wrapping a normal N(μ, σ^2) distribution around the circle. A circular r.v θ is said to have a wrapped normal (WN) distribution if its pdf is $f_W(\theta) = (2\pi)^{-1} + \pi^{-1} \sum_{p=1}^{\infty} \rho^{p^2} \cos[p(\theta - \mu)]$, $0 \le \mu \le 2\pi$, $0 \le \rho \le 1$, where μ and $\rho = \exp\left(\frac{-1}{2}\sigma^2\right)$ are the mean direction and mean resultant length respectively. The value of $\rho = 0$ corresponds to the circular uniform distribution, and as ρ increases to 1, the distribution concentrates increasingly around μ. Stephens (1963) matched the first trigonometric moments of the von Mises and wrapped normal distributions, that is,

$$\rho = \exp\left(\frac{-1}{2}\sigma^2\right) = A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)},$$

establishing that the two have a close relationship, where

$$I_0(\kappa) = (2\pi)^{-1} \int_0^{2\pi} \exp[\kappa \cos(\theta)]d\theta = \sum_{j=0}^{\infty} \frac{1}{(j!)^2} \left(\frac{\kappa^2}{4}\right)^j$$

and

$$I_1(\kappa) = \sum_{r=0}^{\infty} [(r+1)!r!]^{-1} \left(\frac{1}{2}\kappa\right)^{2r+1}$$

are the modified Bessel functions of order zero and order one, respectively. Based on the difficulty in distinguishing the two distributions, Collett and Lewis (1981) concluded that decision on whether to use a von Mises model or a Wrapped Normal model, depends on which of the two is most convenient.

The Wrapped Normal distribution WN (μ, ρ) is obtained by wrapping the N(μ, σ^2) distribution onto the circle, where $\sigma^2 = -2\log\rho$, which implies that, $\rho = \exp\left[\frac{-\sigma^2}{2}\right]$. But for large κ, in particular $\kappa \ge 2$, (Fisher, 1987), VM(μ, κ) is approximately equivalent to N$\left(\mu, \frac{1}{\kappa}\right)$, which in turn is approximately equivalent to

$\mathrm{WN}\left(\mu, \exp\left[\frac{-1}{2\kappa}\right]\right)$. This approximation is very accurate for $\kappa > 10$ (Mardia & Jupp, 2000).

Note $\hat{\sigma}^2 = -2\log A(\kappa)$ and $\hat{\sigma}^2 = \frac{1}{\kappa}$ are the estimates of σ^2 when $\mathrm{VM}(\mu, \kappa)$ is approximated by $\mathrm{WN}(\mu, \rho)$ and $\mathrm{N}\left(\mu, \frac{1}{2\kappa}\right)$ respectively. Figure 1 shows how the WN and N approximations are related for various values of concentration parameter, κ, using the following approximation,

$$A(\kappa) \approx 1 - \frac{1}{2\kappa} - \frac{1}{8\kappa^2} - \frac{1}{8\kappa^3} - \ldots,$$

Jammalamadaka & SenGupta (2001, p. 290).

The circular median is rotationally invariant as shown by Ackermann (1997). Lenth (1981), and, Wehrly and Shine (1981) studied the robustness properties of both the circular mean and median using influence curves, and revealed that the circular mean is quite robust, in contrast to the mean for linear data on the real line. Durcharme and Milasevic (1987), show that in the presence of outliers, the circular median is more efficient than the mean direction. Many authors, including He and Simpson (1992), advocate the use of circular median as an estimate of preferred direction, especially in situations where the data are not from the von Mises distribution.

The Hodges-Lehmann estimator, on the other hand is a compromise between the occasionally non-robust circular mean and the more robust circular median. Unlike the circular median which downweights outliers significantly but is sensitive to rounding and grouping (Wehrly & Shine, 1981), the HL estimate downweights outliers more sparingly and is more robust to rounding and grouping. The circular HL estimator has comparable efficiency to mean and is superior to median; see Otieno and Anderson-Cook (2003a). Other properties of this estimate are explored and compared to those of circular mean and circular median in Otieno and Anderson-Cook (2003a). S-Plus or R functions for computing this estimate are available by request from the authors.

Consider a circular distribution F which is unimodal and symmetric about the unknown direction μ_0. The influence function (IF) for the circular mean direction is given by $IF(\theta) = \frac{\sin(\theta - \mu_0)}{\rho}$, where the mean resultant length is given by $\rho = \exp\left(\frac{-1}{2}\sigma^2\right) = A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)}$. For any given value of ρ, this influence function and its derivative are bounded by $\pm\rho^{-1}$, see Wehrly and Shine (1981). Another result due to Wehrly and Shine (1981) is the influence function of the circular median. Without loss of generality for notational simplicity, assume that $\mu \in [0, \pi]$.

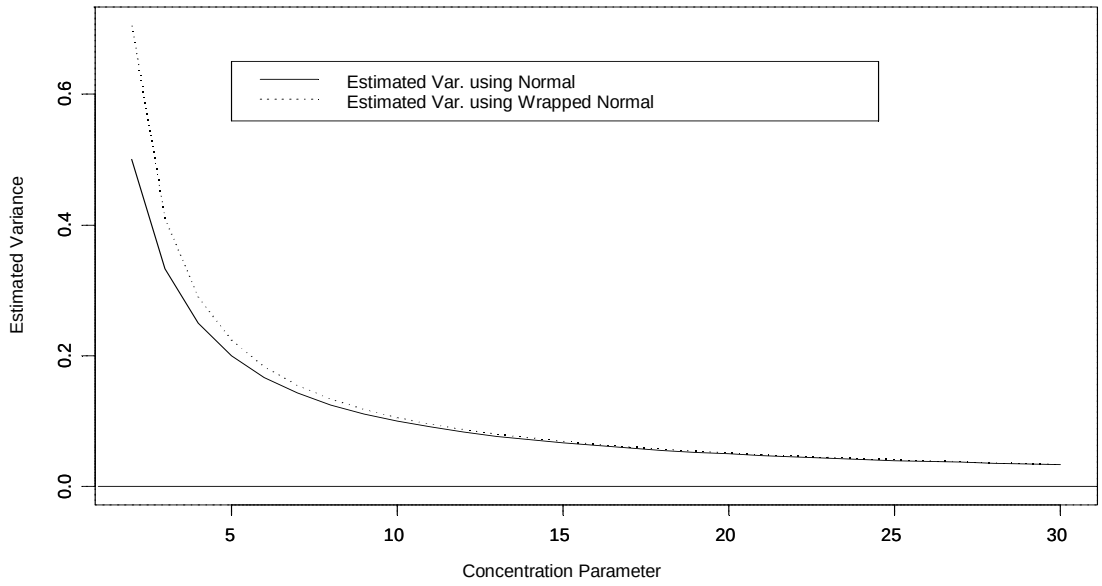
The influence function for the circular median direction is given by

$$IF(\theta) = \frac{\frac{1}{2}\mathrm{sgn}(\theta - \mu_0)}{[f(\mu_0) - f(\mu_0 + \pi)]},$$

$(\mu_0 - \pi < \theta < \mu_0 + \pi)$, where $f(\mu_0)$ is the probability density function of the underlying distribution of the data at the hypothesized mean direction μ_0, and sgn(x) = 1, 0, or -1 as x > 0, x = 0, or x < 0, respectively.

Wehrly and Shine (1981) and Watson (1986) evaluated the robustness of the circular mean via an influence function introduced by Hampel (1968, 1974) and concluded that the estimator is somewhat robust to fixed amounts of contamination and to local shifts, since its influence function is bounded. The influence curve for the circular median, however, has a jump at the antimode. This implies that the circular median is sensitive to rounding or grouping of data (Wehrly & Shine, 1981).

Figure 1: Plot of $\hat{\sigma}^2 = \left[-2\log A\left[\frac{1}{\kappa}\right]\right]$, and $\hat{\sigma}^2 = \left[\frac{1}{\kappa}\right]$ verses Concentration Parameter (κ) for a single observation.



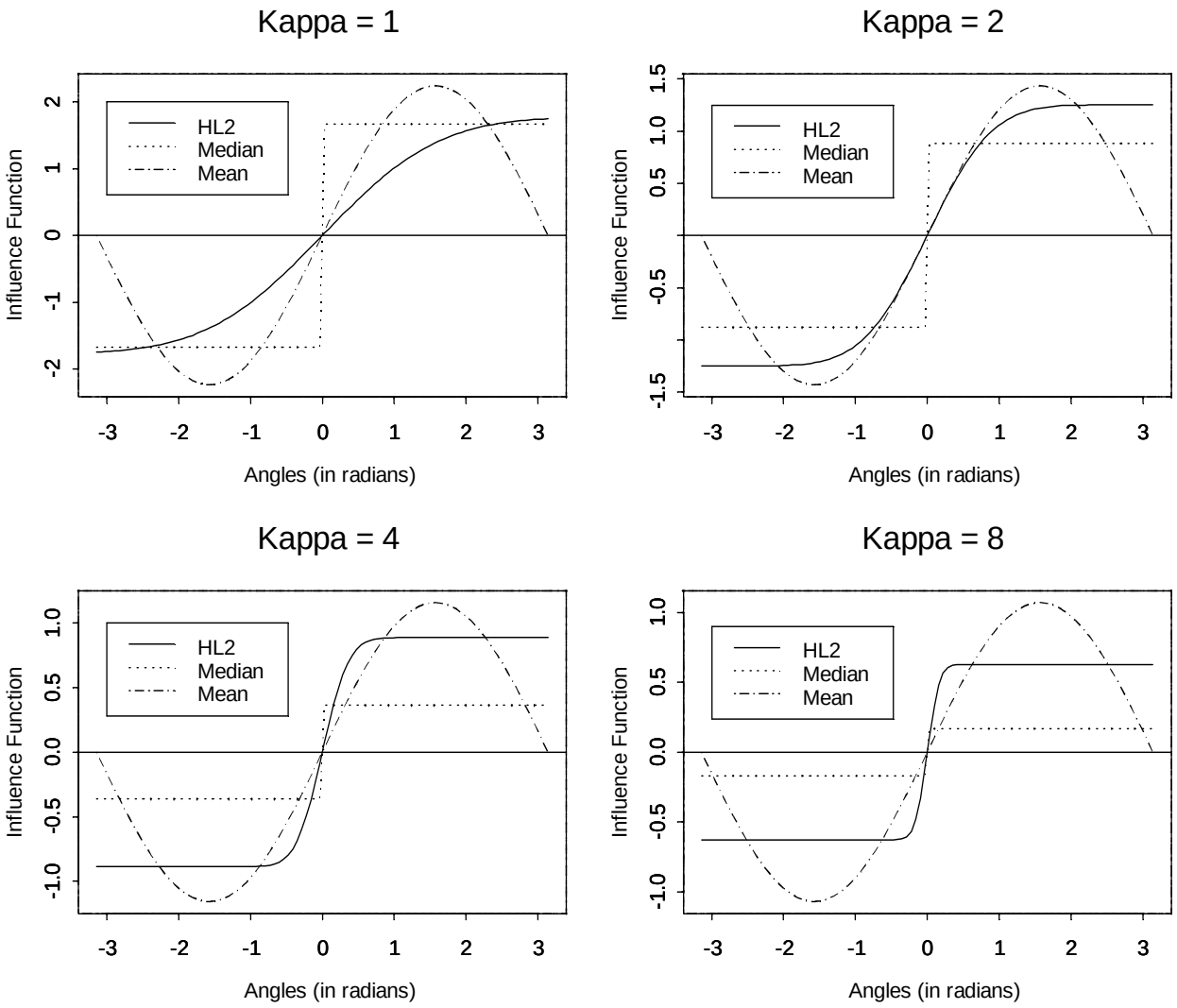
Assume that θ_i and θ_j are iid, with distribution function $F(\theta)$. Let $\Phi = \frac{(\theta_i + \theta_j)}{2}$, $i \le j$. Φ is equivalent to the pairwise circular mean of θ_i and θ_j, Otieno and Anderson-Cook,(2003a). The functional of the circular Hodges-Lehmann estimator $\hat{\theta}^c{}_{HL}$ is the Pseudo-Median Locational functional $F = F^{*-1}\left(\frac{1}{2}\right)$, where $F(\phi) = P(\Phi \le \phi) = \int F(2\phi - \theta)h(\theta)d\theta$, Hettmansperger & McKean (1998, p.3,10-11). For a sample from a von Mises distribution with a limited range of concentrated parameter values, $\kappa \ge 2$, the influence function of the circular HL estimator $\hat{\theta}^c{}_{HL}$ is given by

$$IF(\theta) = \frac{F(\theta) - \frac{1}{2}}{\left(\frac{\kappa}{4\pi}\right)^{\frac{1}{2}}},$$

where F(.) is the cumulative density function of $\theta_1, \ldots, \theta_n$. Note that this influence function is a centered and scaled cdf and is therefore bounded. Note that, it is also discontinuous at the antimode, like the influence function of the circular median.

Figure 2 are plots of the influence functions of the circular mean, circular median and the circular HL estimators for preferred direction for various concentration parameters. The range of the data values is $\frac{-\pi}{2}$ radians to $\frac{\pi}{2}$ radians and 4 dispersion values ranging from κ =1 to 8.

Figure 2: Influence Functions for measures of preferred direction, $(1 \leq \kappa \leq 8)$.



Notice that all the estimators have curves which are bounded. Also, as the data becomes more concentrated (with κ increasing), the influence function of the circular median changes least followed by the circular HL estimator. This is similar to the linear case.

Also, as κ increases, the bound for the influence function for all the three measures decreases, however, overall the bound of the influence function for the mean is largest for angles closest to $\frac{\pi}{2}$ radians from the preferred direction. The maximum influence for the mean

occurs at $\frac{\pi}{2}$ or $\frac{-\pi}{2}$ from the mode for all κ, while for both the median and HL, the maximum occurs uniformly for a range away from the preferred direction. Overall, HL seems like a compromise between the mean and the median.

A Practical Example

Consider the following example of Frog migration data Collett (1980), shown in Figure 3. The data relates the homing ability of the Northern cricket frog, *Acris crepitans*, as studied by Ferguson, et. al.(1967). A number of frogs were collected from mud flats of an abandoned stream meander and taken to a test pen lying to the north of the collection point. After 30 hours enclosure within a dark environmental chamber, 14 of them were released and the directions taken by these frogs recorded (taking 0^0 to be due North), Table 1.

In order to compute the sample mean of these data, consider them as unit vectors, the resultant vector of these 14 unit vectors is obtained by summing them componentwise to get $R = \left(\sum_{i=1}^{n} \cos(\theta_i), \sum_{i=1}^{n} \sin(\theta_i) \right) = (C, S)$, say. The sample circular mean is the angle corresponding to the mean resultant vector $\overline{R} = \frac{R}{n} = \left(\frac{C}{n}, \frac{S}{n} \right) = (\overline{C}, \overline{S})$. That is, the angle corresponding to the mean resultant length $|\overline{R}| = \sqrt{(\overline{C}^2 + \overline{S}^2)}$.For this data the circular mean is -0.977 (124^0), the mean resultant length, $|\overline{R}| = 0.725$, thus, the estimate of the concentration parameter, $\hat{\kappa} = 2.21$ for the best fitting von Mises.(Table A.3, Fisher, 1993, p. 224).

The circular median is -0.816 (133.25^0) and circular Hodges-Lehmann is -0.969 (124.5^0). Using $\hat{\kappa} = 2.21$, Figure 4 gives the influence curves of the mean, median and HL. Note that the measure least influenced by observation x, a presumed outlier, is the circular mean, since x is nearer to the antimode. However, the circular median is influenced most by observations nearest the center of the data followed by HL. The influence of an outlier on the sample circular median is bounded at either a constant positive or a constant negative value, regardless of how far the outlier is from the center of the data. On the other hand, the HL estimator is influenced less by observations near the center, and reflects the presence of the outlier. The influence curve for the circular mean is similar to that of the redescending Φ function (See Andrews et. al., 1972 for details).

Conclusion

Like in the linear case, it is helpful to decide what aspects of the data are of interest. For example, in the case of distributions that are not symmetric or have outliers, like in the case of the Frog migration data, the circular mean and circular median are measuring different characteristics of the data. Hence one needs to choose which aspect of the data is of most interest. For data that are close to uniformly distributed or have rounding or grouping, it is wise to avoid the median since its estimate is prone to undesirable jumps. Either of the other two measures perform similarly. For data spread on a smaller fraction of the circle, with a natural break in the data, the median is least sensitive to outliers. The mean is typically most responsive to outliers, while HL gives some, but not too much weight to outliers.

Overall, the circular HL is a good compromise between circular mean and circular median, like its counterpart for linear data. The HL estimator is less robust to outliers compared to the median, however it is an efficient alternative, since it has a smaller circular variance, Otieno and Anderson-Cook, (2003a). The HL estimator also provides a robust alternative to the mean especially in situations where the model of choice of circular data (the von Mises distribution) is in doubt. Overall, the circular HL estimate is a solid alternative to the established circular mean and circular median with some of the desirable features of each.

Figure 3:The Orientation of 14 Northern Cricket Frogs

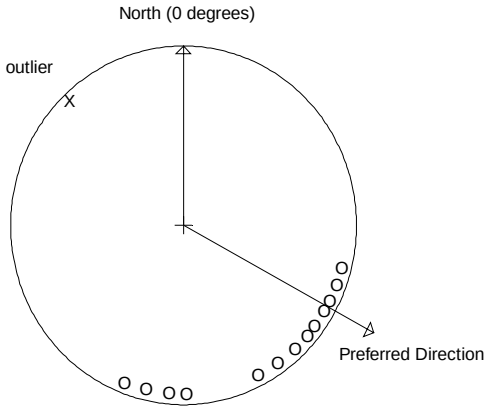
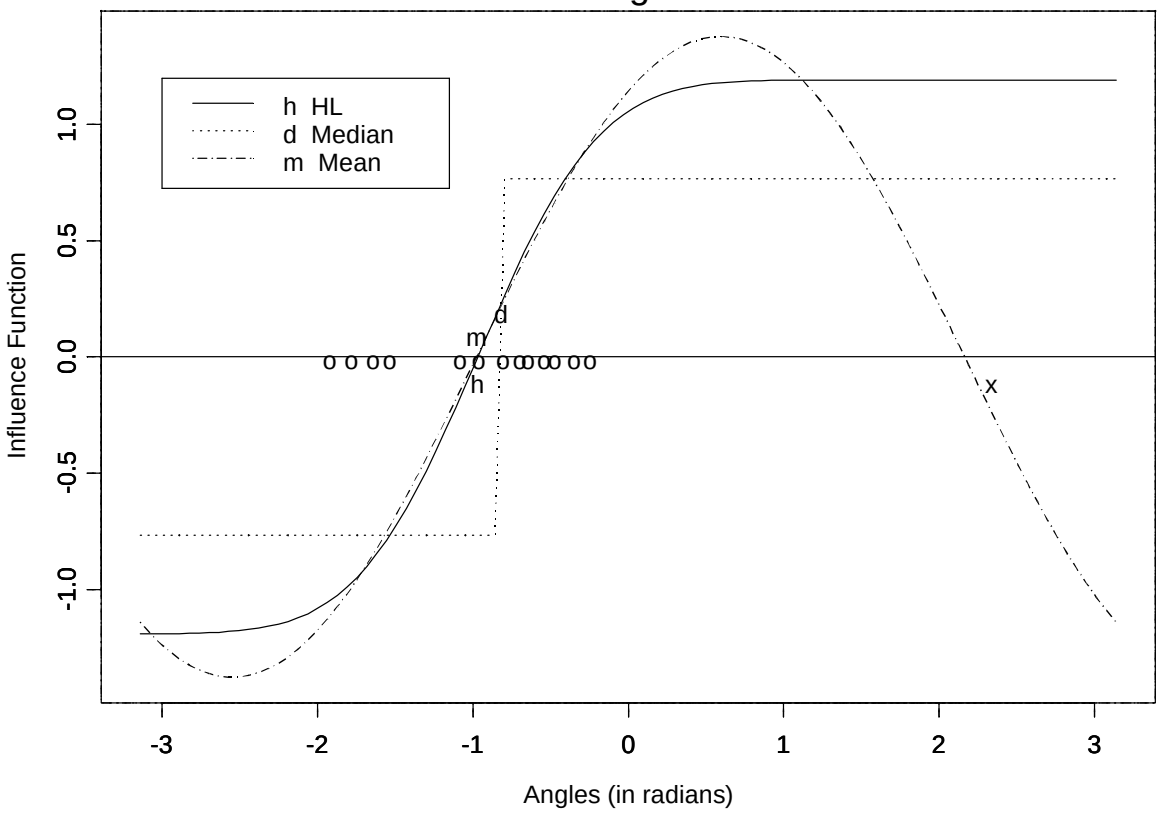


Figure 4: Influence curves for the three measures for data with a single outlier



References

Ackermann, H. (1997). A note on circular nonparametrical classification. *Biometrical Journal*, *5*, 577-587.

Andrews, D. R., Bickel, P. J., Hampel,F. R., Huber, P. J., Rogers, W. H., & Tukey, J. W. (1972). *Robust estimates of location: Survey and advances.* New Jersey: Princeton University Press.

Collett, D. (1980). Outliers in circular data. *Applied Statistics*, *29*, 50-57.

Collett, D., & Lewis, T. (1981). Discriminating between the von Mises and wrapped normal distributions. *Australian Journal of Statistics*, 23, 73-79.

Ferguson, D. E., Landreth, H. F., & McKeown, J. P. (1967). Sun compass orientation of northern cricket frog, Acris crepitans. *Animal Behaviour*, *15*, 45-53.

Fisher, N. I. (1993). *Statistical analysis of circular data.* Cambridge University Press.

Ko, D., & Guttorp P. (1988). Robustness of estimators for directional data. *Annals of Statistics*, *16*, 609-618.

Hettmansperger, T. P. (1984). *Statistical inference based on ranks*. New York: Wiley.

Hettmansperger, T. P., & McKean, J. W. (1998). *Robust nonparametric statistical methods*. New York: Wiley.

Jammalamadaka, S. R., & SenGupta, A. (2001). *Topics in circular statistics, world scientific*. New Jersey.

Mardia, K. V. (1972) *Statistics of directional data.* London: Academic Press.

Mardia, K. V., & Jupp, P. E. (2000). *Directional statistics*. Chichester: Wiley.

Otieno, B. S. (2002) *An alternative estimate of preferred direction for circular data*. Ph.D Thesis., Department of Statistics, Virginia Tech. Blacksburg: VA.

Otieno, B. S., & Anderson-Cook, C. M. (2003a). Hodges-Lehmann estimator of preferred direction for circular. *Virginia Tech Department of Statistics Technical Report*, 03-3.

Otieno, B. S., & Anderson-Cook, C. M. (2003b). A More efficient way of obtaining a unique median estimate for circular data. *Journal of Modern Applied Statistical Methods*, *3*, 334-335.

Rao, J. S. (1984) Nonparametric methods in directional data. In P. R. Krishnaiah and P. K. Sen. (Eds.), *Handbook of Statistics*, 4, pp. 755-770. Amsterdam: Elsevier Science Publishers.

Stephens, M. A. (1963). Random walk on a circle. *Biometrika*, *50*, 385-390.

Watson, G. S. (1986). Some estimation theory on the sphere. *Annals of the Institute of Statistical Mathematics*, 38, 263-275.

Wehrly, T., & Shine, E. P. (1981). Influence curves of estimates for directional data. *Biometrika*, *68*, 334-335.