Book of the Short Papers

Editors: Francesco Maria Chelli, Mariateresa Ciommi, Salvatore Ingrassia, Francesca Mariani, Maria Cristina Recchioni

SIS Società Italiana di Statistica

UNIVERSITÀ POLITECNICA DELLE MARCHE

Dipartimento di Scienze Economiche e Sociali DISES

LIUC Università Cattaneo

LIUC Università Cattaneo | BUSINESS ANALYTICS AND DATA SCIENCE HUB

Pearson

CHAIRS

Salvatore Ingrassia (Chair of the Program Committee) - *Università degli Studi di Catania*
Maria Cristina Recchioni (Chair of the Local Organizing Committee) - *Università Politecnica delle Marche*

PROGRAM COMMITTEE

Salvatore Ingrassia (Chair), Elena Ambrosetti, Antonio Balzanella, Matilde Bini, Annalisa Busetta, Fabio Centofanti, Francesco M. Chelli, Simone Di Zio, Sabrina Giordano, Rosaria Ignaccolo, Filomena Maggino, Stefania Mignani, Lucia Paci, Monica Palma, Emilia Rocco.

LOCAL ORGANIZING COMMITTEE

Maria Cristina Recchioni (Chair), Chiara Capogrossi, Mariateresa Ciommi, Barbara Ermini, Chiara Gigliarano, Riccardo Lucchetti, Francesca Mariani, Gloria Polinesi, Giuseppe Ricciardo Lamonica, Barbara Zagaglia.

ORGANIZERS OF INVITED SESSIONS

Pierfrancesco Alaimo Di Loro, Laura Anderlucci, Luigi Augugliaro, Ilaria Benedetti, Rossella Berni, Mario Bolzan, Silvia Cagnone, Michela Cameletti, Federico Camerlenghi, Gabriella Campolo, Christian Capezza, Carlo Cavicchia, Mariateresa Ciommi, Guido Consonni, Giuseppe Ricciardo Lamonica, Regina Liu, Daniela Marella, Francesca Mariani, Matteo Mazziotta, Stefano Mazzuco, Raya Muttarak, Livia Elisa Ortensi, Edoardo Otranto, Ilaria Prosdocimi, Pasquale Sarnacchiaro, Manuela Stranges, Claudia Tarantola, Isabella Sulis, Roberta Varriale, Rosanna Verde.

FURTHER PEPOPLE OF LOCAL ORGANIZING COMMITTEE

Elisa D'Adamo, Christian Ferretti, Giada Gabbianelli, Elvina Merkaj, Luca Pedini, Alessandro Pionati, Marco Tedeschi, Francesco Valentini, Rostand Arland Yebetchou Tchounkeu

Technical support: Matteo Mercuri, Maila Ragni, Daniele Ripanti

PUBLISHED BY PEARSON
WWW.PEARSON.COM
ISBN 9788891935618AAVV

Anomaly Detection in Circular Data

Houyem Demni[a] and Giovanni C. Porzio[a]

[a]Department of Economics and Law, University of Cassino and Southern Lazio;
houyem.demni@unicas.it, porzio@unicas.it

Abstract

Circular data arise as directions, rotations, axes, clock, or calendar measurements. Applications are found in industry, envirometrics, Earth sciences and many other fields. Detecting outliers is an important problem that has been studied in several research areas. In this study, an outlier identification procedure for circular data is suggested. The proposed method is based on robust estimates of distribution parameters on the circle and it is illustrated through two real data examples.

Keywords: angles, directions, Ko estimator, outliers, robust statistics.

1. Introduction

A circular observation lies on the circumference of the unit circle and it can be described in polar coordinates by an angle $\phi \in [-\pi, \pi)$ or $[0, 2\pi)$ measured in a specified direction from a specified origin, as well as in Cartesian coordinates through the vector $x = (\cos\phi, \sin\phi)^T$ for which $||x|| = 1$. Circular data arise in many fields such as in Earth sciences (5), biology (20), bioinformatics (16) and also in industry (11). Books covering many aspects of circular data are available within the literature (15; 23).

When dealing with circular data, as with any kind of data analysis, outlying observations or anomalies may influence the main findings and conclusions. They can also reveal unexpected patterns in the data.

Outliers can be defined as observations that are different from the majority. These outlying observations may occur due to copying or recording errors, they could have been recorded under exceptional circumstances, or they simply come from another population.

Detecting these anomalous cases can be thus essential. However, numerous difficulties can arise while performing this task. In practice, as it will be discussed shortly, we found that the available techniques may be not as effective as they should be. Particularly, outliers may not be detected, a notorious effect called masking, or some good observations might be flagged as outliers (which is known as the swamping effect). To avoid these effects, a potentially useful approach is to rely on robust statistical procedures, and this work is aimed at investigating this perspective.

The paper is organized as follows. Section 2. provides a review on outlier detection techniques on the circle, while Section 3. describes the robust anomaly detection technique. Finally, in Section 4, two real data examples are used in order to illustrate the proposed methodology.

2. Outlier Detection on the Circle

Within the literature, several tools have been considered to detect outliers in circular data. One option is to detect outliers by deletion. That is, one or more points are deleted, the analysis is performed without

them, the deleted points are then somehow compared with the obtained results. Within this context, many techniques have been made available. For instance, a statistic that identifies an observation as an outlier if it appears as the most influential observation on the mean resultant length has been proposed (14). Four tests of discordancy for outlier detection have been described and compared in (6). These techniques can be only used for small sample sizes and to detect a single outlier. A discussion on outlier detection on the circle has been also provided in textbooks (7; 15) where the proposal of (6; 14) has been considered.

An outlier detection rule based on the locally most powerful invariant statistic and the likelihood ratio test has been introduced in (21) under the assumption that the data follow a von Mises distribution. They compared their proposed method with the ones in (6; 14). However, their method relies on assuming that the concentration parameter is known.

Unfortunately, outlier detection techniques by deletion suffer from the masking effect. That is, an outlier is undetected because of the presence of another adjacent anomalous observation.

More recently, a series of new statistical tests for anomaly detection in circular data, based on a circular distance (3; 12), the sums of these distances (2) and on the spacings theory (17) have been introduced and compared with existing techniques. Nevertheless, each of these techniques has some limits. The procedures in (2; 3) are able to detect only single outliers, while the cutoff value for the one in (12; 17) is obtained through simulations under a specific data model. Additionally, the detection rule in (17) imposes that multiple outliers are well separated from the rest of the data.

Other authors discussed how to identify outliers in multivariate directional settings (i.e., when data lie on a sphere or on a torus) (1; 8; 24). Although these methodologies can be adapted to the circular case, no specific study is available along this direction.

Alternatively, robust statistical techniques can be used. However, this concept have been only considered in (4) or within the context of circular regression (19). In (4), the weighted likelihood and minimum disparity methods are extended to the circular case under the von Mises distribution assumption. Their proposal is rather complex to be applied and it strongly depends on the choice of a bandwidth and of a certain α parameter.

3. Robust Anomaly Detection

Anomaly detection is a task strongly related to the idea of robust statistics. Outliers can be detected by fitting the majority of the data and flagging as potential outliers the observations that deviate from it (22).

Robust procedures assume that the majority of the data (that are supposed to be clean) follows a specific probability distribution (9). For instance, for data on a line, the data are assumed to follow the Normal probability density function with unknown mean and standard deviation. Under this assumption, the location and dispersion parameters of the distribution are estimated in a robust way, and a cutoff threshold is identified in order to recognize and discard potential outliers. As cutoff, the quantile of the assumed distribution is typically considered.

For Normal data, thus, an observation x_i will be flagged as outlier if

$$\frac{x_i - \hat{\mu}}{\hat{\sigma}} < \Phi^{-1}(1 - \alpha/2),$$

where $\hat{\mu}$ and $\hat{\sigma}$ are some robust estimates of the corresponding parameters, and Φ is the standard Normal cumulative distribution function (cdf).

Within the circular domain, we apply this same procedure and we assume data come from the von Mises distribution. The von Mises distribution is the most used distribution to model circular data, and its circular density is given by:

$$h(\phi; \mu, \kappa) := \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\phi - \mu)), \tag{1}$$

with I_0 the modified Bessel function of the first kind and of order 0, and where μ is a location parameter and $\kappa \geq 0$ is the concentration parameter. For $\kappa = 0$, the distribution reduces to the uniform distribution

on the circle. When $\kappa > 0$, the distribution is symmetric around μ, which is both the directional mean and median of the distribution.

In our setting, we also assume $\kappa > 0$. This is because (a) under uniformity on the circle it would be odd to find points that "deviate from the majority of the data", and (b) in such a case the parameter μ is undefined.

Under this model, robust anomaly detection will be performed by first robustly estimating μ and κ. Then, the (shortest arc) distance of each observed angle ϕ_i from the estimated μ will be computed, and then compared with a cutoff value $c_\alpha(\kappa), \alpha > 0$. The value of α will be set by the analyst, keeping in mind that it represents the expected proportion of the points that will be flagged as outliers while actually they are not.

The cutoff value $c_\alpha(\kappa)$ will be the $1 - \alpha/2$ quantile direction of a von Mises distribution centered in $\mu = 0$. It will be thus obtained by solving the equation: $\int_0^{c_\alpha(\kappa)} \frac{1}{2\pi I_0(\kappa)} \exp(\kappa \cos(\phi))d\phi = (1 - \alpha/2)$.

Hence, in practice, a circular observation ϕ_i will be flagged as potential outlier/anomalous data if

$$d(\phi_i, \hat{\mu}) > c_\alpha(\hat{\kappa}), \tag{2}$$

where $d(\phi_i, \phi_j) := \pi - |\pi - |\phi_i - \phi_j||$ is the length of the shortest arc joining ϕ_i with ϕ_j.

Robust estimators of μ and κ must be adopted in Equation 2 in order to get an effective anomaly detection rule. This will guarantee protection against the masking effect, while the level of swapping will be controlled by the chosen value of α.

Within this work, as robust estimators of μ and κ, we suggest to use the Fisher circular median and the simple concentration estimator discussed in (10), respectively. The first is defined as the point minimizing the shortest arc distances of the sampled observations from it. That is, let $C = \{\phi_1, .., \phi_i, .., \phi_n\}$ be a circular data set. Its Fisher median is given by:

$$\hat{\mu} := \arg\min_{\eta \in \mathcal{S}} \sum_{i=1}^{n} d(\phi_i, \eta)). \tag{3}$$

The robust estimator of the concentration parameter described in (10) is instead given by

$$\hat{\kappa} = (\Phi^{-1}(0.75)/CMAD(C))^2, \tag{4}$$

where $CMAD(C)$ is the circular median absolute deviation of the set C, this latter being the median of the shortest arc distances of the observed values ϕ_i from $\hat{\mu}$.

At the end, as a peculiar property of data on the circle, we note that the minimization problem in Equation 3 can result in a disconnected set of values (if the set is connected, the median will be given by its central point). Should this unlikely event occur, a different robust estimator of the location parameter μ must be adopted (e.g. the circular trimmed mean estimator).

4. Illustrative Examples

For illustrative purposes, the anomaly detection procedure proposed in Section 3. is here applied to two real data sets. The first is the well-known Sardinian sea stars while the second is related to an industrial application, and it considers some wind directions.

4.1 Sea stars

The Sea stars data was provided by (7) and it is available within the library *circular* in *R*. It refers to the resultant directions in degrees moved by 22 Sardinian sea stars over a period of 11 days after their displacement from their natural habitat. We transform the given angles from degrees to radians, and we consider that 0 radians is the North pole and the rotation is clockwise.

The data are plotted in Figure 1. According to (7), the 13th and 14th observations (2.565 and 5.201 radians) are outliers. In fact, these values emerge as far-out values with respect to the the majority of the

data. The sample circular median of the data is $\hat{\mu} = 0.040$ radians and it is shown by the black arrow in Figure 1 (left and right panels) while the corresponding estimator of the concentration parameter $\hat{\kappa} = 6.942$. Then, we compute the cutoff threshold at a level $\alpha = 0.01$ (which turns out to be equal to 0.936 radians), and the shortest arc distances between each data point and the circular median. By comparing the computed distances with the threshold, observations 13 and 14 are flagged as outliers (Figure 1, right panel, highlighted in red).

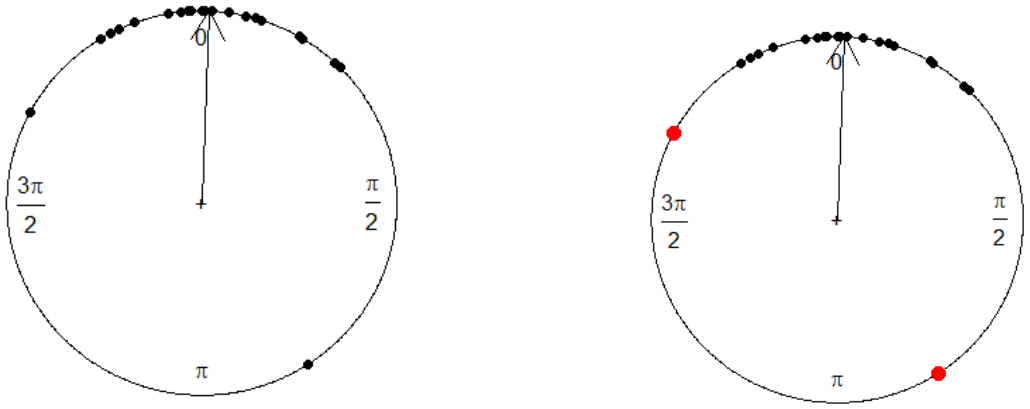

Figure 1: Resultant directions moved by Sardinian sea stars over a period of 11 days after displacement from their natural habitat. Raw data circular plots: the black arrow shows the sample circular median direction (left), outliers are flagged in red (right).

4.2 Wind directions

The modeling and the monitoring of wind directions play an important role in the industry of wind power generation (18). The data comes from the recording of wind directions from the meteorological station at "Col de la Roa" in the Italian Alps via data-logger every 15 minutes. Daily recorded wind directions between 3:00 am and 4:00 am inclusive from January 29, 2001 to March 31, 2001 are considered. Accordingly, there are five directions recorded every day leading to a total of 310 measurements (in radians). The data are also available within the *R* package *circular*.

These wind directions have a sample circular median $\hat{\mu} = 0.165$ radians and an estimate of the concentration $\hat{\kappa} = 3.848$. The associated cutoff value at $\alpha = 0.01$ is given by 1.351 radians. The wind directions are depicted in Figure 2 (left panel) and their circular median is drawn by the black arrow. By evaluating the shortest arc distances between each point and the median, and comparing them to the threshold, outliers are flagged (Figure 2, right panel). We found sub-populations of outliers located around the East-Southeast, Southeast, South and West directions.

The same data example was considered in (4), where the presence of sub-populations of outliers located around the East-Southeast, Southeast and South directions was visually inferred by means of a non parametric density estimator.

Acknowledgments This work has been partially funded by the BiBiNet project (grant H35F21000430002) within the POR-Lazio FESR 2014-2020.

References

[1] Abuzaid, A. H.: Identifying density-based local outliers in medical multivariate circular data. Stat. Med. **39(21)**, 2793-2798 (2020).

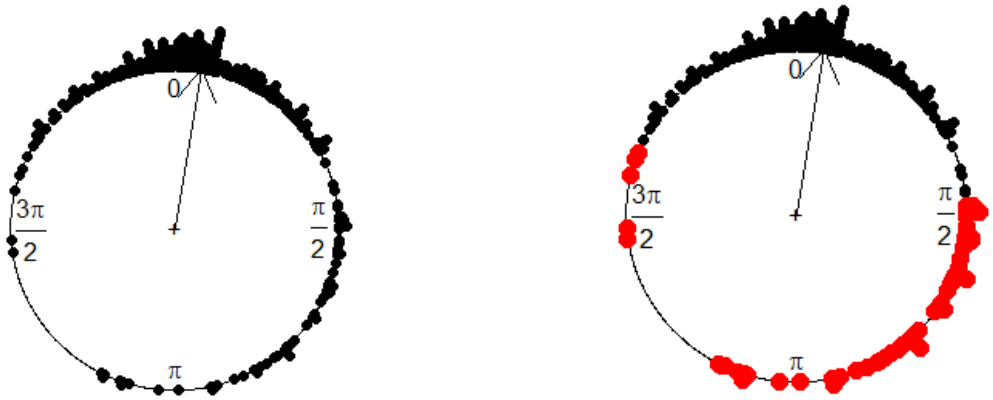

Figure 2: Daily recorded wind directions. Raw data circular plots: the black arrow shows the sample circular median direction (left), outliers are flagged in red (right).

[2] Abuzaid, A. H., Mohamed, I. B., and Hussin, A. G.: A new test of discordancy in circular data. Commun. Stat. Simul. Comput. **38(4)**, 682-691 (2009) doi: 10.1080/03610910802627048.

[3] Abuzaid, A. H., Hussin, A. G., Rambli, A., and Mohamed, I.: Statistics for a new test of discordance in circular data. Commun. Stat. Simul. Comput. **41(10)**, 1882-1890 (2012) doi:10.1080/03610918.2011.624239.

[4] Agostinelli, C.: Robust estimation for circular data. Comput. Stat. Data. Anal. **51(12)**, 5867-5875 (2007).

[5] Cabella, P. and Marinucci, D.: Statistical challenges in the analysis of cosmic microwave background radiation. Ann. Appl. Stat. **3(1)**, 61-95 (2009) doi:10.1214/08-aoas190.

[6] Collett, D.: Outliers in circular data. J. R. Stat. Soc. C-Appl. **29(1)**, 50-57 (1980).

[7] Fisher, N. I.: Statistical analysis of circular data. Cambridge, UK: Cambridge University Press. (1993) doi: 10.1017/cbo9780511564345.

[8] Greco, L., Saraceno, G., and Agostinelli, C.: Robust fitting of a wrapped normal model to multivariate circular data and outlier detection. Stats. **4(2)**, 454-471 (2021).

[9] Hubert, M., and Van der Veeken, S.: Outlier detection for skewed data. J. Chemom. **22(3-4)**, 235-246 (2008) doi: 10.1016/j.simpat.2018.05.010.

[10] Ko, D.: Robust estimation of the concentration parameter of the von Mises-Fisher distribution. Ann. Stat. **20(2)** 917-928 (1992).

[11] Lima–Filho, L. M., Bayer, F. M., and da Silva, A. M.: Control chart to monitor circular data. Qual. Reliab. Eng. **37(3)**, 966-983 (2021).

[12] Mahmood, E. A., Rana, S., Midi, H., and Hussin, A. G.: Detection of outliers in univariate circular data using robust circular distance. J. Mod. Appl. Stat. Methods. **16(2)** 22 (2017) doi:10.22237/jmasm/1509495720.

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

[14] Mardia, K. V.: Statistics of directional data. J. R. Stat. Soc. Series B Stat. Methodol. **37(3)**, 349-371 (1975).

[15] Mardia, K. V., and Jupp, P. E.: Directional statistics. Chichester, UK: John Wiley and Sons Ltd. (2000) doi: 10.1002/9780470316979.

[16] Mardia, K. V., Hughes, G., Taylor, C. C., and Singh, H.: A multivariate von Mises distribution with applications to bioinformatics. Can. J. Stat. **36(1)**, 99-109 (2008) doi:10.1002/cjs.5550360110

[17] Mohamed, I. B., Rambli, A., Khaliddin, N., and Ibrahim, A. I. N.: A new discordancy test in circular data using spacings theory. Commun. Stat. Simul. Comput. **45(8)**, 2904-2916 (2016).

[18] Koivisto M., Ekström J., Mellin I., Millar J., Lehtonen M.: Statistical wind direction modeling for

the analysis of large scale wind power generation. Wind. Energy **20(4)**, 677-694 (2017).
[19] Rana, S., Mahmood, E. A., Midi, H., and Hussin, A. G.: Robust detection of outliers in both response and explanatory variables of the simple circular regression model. Malays. J. Math. Sci. **10(3)**, 399-414 (2016).
[20] Ranalli, M., and Maruotti, A.: Model-based clustering for noisy longitudinal circular data, with application to animal movement. Environmetrics **31(2)**, e2572 (2020).
[21] Rao, J. S., and Sengupta, A.: Topics in circular statistics. World Scientific Press, Singapore, **10**, 4031 (2001) doi: 10.1142/4031.
[22] Rousseeuw, P. J., and Hubert, M.: Anomaly detection by robust statistics. Wiley. Interdiscip. Rev. Data. Min. Knowl Discov. **8(2)**, e1236 (2018) doi: 10.1002/widm.1236.
[23] Pewsey, A., Neuhäuser, M., and Ruxton, G. D.: Circular statistics in R. Oxford University Press, UK (2013).
[24] Sau, M. F., and Rodriguez, D.: Minimum distance method for directional data and outlier detection. Adv. Data. Anal. Classif. **12**, 587-603 (2018).