

# Outlier Classification and Exclusion in Circular Data

A Supplement to “*Working with Circular Data: A Tutorial for Cognitive and Behavioral Research*”

Stella Wernicke<sup>1\*</sup>, Julie de Falco<sup>1</sup>, Zahara Gironés<sup>1,2,3</sup>, Chris Jungerius<sup>1</sup>,  
Ivan Tomić<sup>4</sup>, Adam Triabhall<sup>1</sup>, Xiaolu Wang<sup>1,5</sup>, and Paul M Bays<sup>1</sup>

<sup>1</sup>Department of Psychology, University of Cambridge, United Kingdom

<sup>2</sup>Department of Engineering, University of Cambridge, United Kingdom

<sup>3</sup>Department of Clinical Neuroscience, University of Cambridge, United Kingdom

<sup>4</sup>Department of Psychology, University of Zagreb, Croatia

<sup>5</sup>Department of Psychology, Sun Yat-sen University, China

\*Corresponding author: [saiw2@cam.ac.uk](mailto:saiw2@cam.ac.uk)

As in Euclidean statistics, an outlier in circular data is an observation that deviates markedly from others. In behavioral and cognitive research, such values may reflect measurement errors, misunderstandings of the task, lapses in attention, or meaningful extreme values. In circular statistics, detecting outliers requires special care because the ordering of the data is not unique and minimum and maximum values are not defined. Outliers should be excluded sparingly to avoid biasing the results; in many cases, no exclusions will be necessary at all. In this supplemental material, we briefly survey some methods for detecting outliers, from formal statistical tests to more algorithmic classification techniques.

One approach to detecting outliers in circular data is to use *statistical goodness-of-fit tests* that check whether the sample is drawn from an expected distribution. These tests can be appropriate if researchers have clear expectations about the data’s distribution based on previous empirical findings or theoretical considerations. If the data deviate significantly from an expected distribution, this *may* indicate the presence of one or more outliers. Landler et al. (2018) give an overview of a range of commonly used tests; Landler et al. (2020) discuss how to modify them for naturally discrete or rounded data. While a significant result suggests the presence of outliers, statistical tests do not pinpoint which observations might be outliers.

Methods based on *descriptive statistics* (see the [Descriptive Statistics of Circular Data](#) section) identify outliers by analyzing how far each observation lies from a central value. In Euclidean statistics, it is an established approach to base the threshold for outliers on z-scores. For example, outliers are sometimes defined as values that deviate from the mean by at least three standard deviations. Analogously, extreme values in circular space can be excluded using a cutoff of a maximum deviation from the circular mean in terms of circular standard deviations. However, outlier values can distort the circular mean and inflate the circular standard deviation. Therefore, alternative outlier classification methods use measures that

are more robust against influences from outliers. The *median absolute deviation* is a robust measure of spread that can be used together with the circular median to identify outliers in circular data. It is calculated by computing the angular distance of each observation from the circular median, and then calculating the Euclidean median of those distances. Outliers can now be defined as values that deviate from the circular median by the median absolute deviation multiplied by some factor (e.g., three). While methods based on distance and deviation provide an intuitive approach toward identifying concrete outlier candidates, they are sometimes criticized for being exploratory and subjective.

Another approach to detecting outliers is to identify *patterns in the total data*, rather than comparing individual data points to the overall summary. These methods usually employ *clustering and density-based algorithms*. Clustering algorithms structure data into subsets, aiming for high similarity between data points in one cluster and low similarity between data points of different clusters. If most of the data form one large cluster and a few points form one or several tiny separate clusters, those few points can be deemed outliers. There are circular versions of many established clustering algorithms that group data points based on angular proximity, including circular versions of k-means (Charalampidis, 2005), fuzzy c-means (Kesemen et al., 2016), and mean-shift (Zhang & Chen, 2021). While these methods allow for identifying multiple concrete outlier candidates, their results depend on the choice of the algorithm and hyperparameters.

## References

- Charalampidis, D. (2005). A modified k-means algorithm for circular invariant clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(12), 1856–1865. <https://doi.org/10.1109/TPAMI.2005.230>

- Kesemen, O., Tezel, Ö., & Özkul, E. (2016). Fuzzy c-means clustering algorithm for directional data (fcm4dd). *Expert Systems with Applications*, 58, 76–82. <https://doi.org/https://doi.org/10.1016/j.eswa.2016.03.034>
- Landler, L., Ruxton, G. D., & Malkemper, E. P. (2018). Circular data in biology: Advice for effectively implementing statistical procedures. *Behavioral Ecology and Sociobiology*, 72(8), 128. <https://doi.org/10.1007/s00265-018-2538-y>
- Landler, L., Ruxton, G. D., & Malkemper, E. P. (2020). Grouped circular data in biology: Advice for effectively implementing statistical procedures. *Behavioral Ecology and Sociobiology*, 74(8), 100. <https://doi.org/10.1007/s00265-020-02881-6>
- Zhang, Y., & Chen, Y.-C. (2021). Kernel smoothing, mean shift, and their learning theory with directional data. *Journal of Machine Learning Research*, 22(154), 1–92. <http://jmlr.org/papers/v22/20-1194.html>