# Improvements to the Massive Unsupervised Outlier Detection (MUOD) Algorithm*

Oluwasegun Ojo †1,3, Antonio Fernández Anta1, and Rosa E. Lillo2

1IMDEA Networks Institute, Madrid, Spain
2UC3M-Santander Big Data Institute, Universidad Carlos III de Madrid, Spain
3Universidad Carlos III de Madrid, Spain

We present improvements to the Massive Unsupervised Outlier Detection (MUOD) algorithm, a scalable and unsupervised outlier detection method, especially useful for identifying outliers for functional data. MUOD identifies different types of outliers in samples of curves including shape, magnitude and amplitude outliers. This is done by computing for each curve three indices, which measure outlyingness in terms of shape, magnitude and amplitude relative to the other curves. These indices are then sorted and observations with extremely high indices are labelled as outliers. To further improve the scalability MUOD, we introduce "fastMUOD", a fast implementation of MUOD which uses the component-wise or the $L_1-$median in the computation of the indices instead of using the whole observation. We also present "semi-fastMUOD", which uses a sample of the observations in the computation of the indices. As further improvements to MUOD, we discuss a new method for identifying extreme indices which entails the use of a classical boxplot or its adjusted version for skewed distributions. We analyse the performance of the proposed improvements using real and simulated data, and show that outlier detection accuracy is not compromised even with the gains in scalability.

## References

[1] Azcorra, A., et al. (2018). Unsupervised scalable statistical method for identifying influential users in online social networks. *Scientific Reports*, **8**, 6988.

[2] Hubert, M., Rousseeuw, P. J., and Segaert, P. (2015). Multivariate functional outlier detection. *Statistical Methods and Applications*, **24**, 177-202.

[3] Sun, Y., and Genton, M. G., (2011). Functional boxplots. *Journal of Computational and Graphical Statistics*, **20**, 316-334.

[4] Hubert, M., Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*, **52**, 5186-5201.

[5] Febrero, M., Galeano, P., and González-Manteiga, W. (2008). Outlier detection in functional data by depth measures, with application to identify abnormal NOX levels. *Environmetrics*, **19**, 331-345.

*Partially supported by the Spanish MICIU grant DiscoEdge (TIN2017-88749-R), the Comunidad de Madrid grant EdgeData-CM (P2018/TCS-4499), and the NSF of China grant 61520106005.

†Corresponding author. Email: oluwasegun.ojo@imdea.org.