Geometric and Topological Foundations of Big Data Analytics

Fionn Murtagh (1, 2), Pedro Contreras (3)

(1) Department of Computing and Mathematics, University of Derby. (2) Department of Computing, Goldsmiths University of London. (3) Thinking Safe Ltd., Egham. Email: fmurtagh@acm.org, pedro.contreras@acm.org

In Murtagh and Contreras (2015) we set out the foundations of clustering based on seriation, and using random projection. We relate this to stochastic approximation (cf. Benzécri, 1982, discussing the analysis of an infinite, i.e. unbounded, set of observations crossed by 1000 attributes); to power iteration clustering; and to how hierarchical clustering can be perfectly scaled in one dimension. Such are the theoretical underpinnings of our work in astronomy, chemoinformatics, and other applications (see Contreras and Murtagh, 2012).

The potential is considerable in regard to social media analytics and other applications (Murtagh et al., 2015). A most interesting viewpoint of Keiding and Louis (2016) is that statistical sampling can benefit from Big Data calibration. In our contribution to the discussion following the presentation of this paper, we noted how the need to bridge sampled data and calibrating Big Data can be addressed, for the data analyst and for the application specialist, through the geometry and the topology of information and data. (Cf. also Murtagh, 2013.)

References

1. J.P. Benzécri, "L'approximation stochastique en analyse des correspondances", *Les Cahiers de l'Analyse des Données*, 7, 387–394, 1982.
2. P. Contreras and F. Murtagh, "Fast, linear time hierarchical clustering using the Baire metric", *Journal of Classification*, 29, 118–143, 2012.
3. N. Keiding and T.A. Louis (2016), "Perils and potentials of self-selected entry to epidemiological studies and surveys", *Journal of the Royal Statistical Society A*, 179, Part 2, 1–28, 2016
4. F. Murtagh, "The new science of complex systems through ultrametric analysis: Application to search and discovery, to narrative and to thinking", *Journal of p-Adic Numbers, Ultrametric Analysis and Applications*, vol 5, no. 4, 326-337, 2013.
5. F. Murtagh and P. Contreras, "Clustering through high-dimensional data scaling: applications and implementations", ECDA 2015 (European Conference on Data Analysis) Conference presentation, September 2015. *Archives of Data Science*, submitted, 2015.
6. F. Murtagh, M. Pianosi, R. Bull, "Semantic mapping of discourse and activity, using Habermas's Theory of Communicative Action to analyze process", *Quality and Quantity*, in press.