Addressing New Challenges, including Ethics, of Big Data Analytics, with the Correspondence Analysis and Geometric Data Analysis Platform

Fionn Murtagh, University of Derby, and Goldsmiths University of London.
fmurtagh@acm.org

Abstract

Challenges to be addressed in Big Data analytics are in the comprehensive survey of Keiding and Louis (2016) that sets out new contemporary issues of sampling and population distribution estimation. As I note in the commentary section of this publication, while "Representativity should be avoided", the bridge between the data that is analyzed, and the calibrating Big Data, is well addressed by the geometry and topology of data. Those form the link between sampled data and the greater cosmos.

Our analytics can well achieve the "rehabilitation of individuals" in analytics, as referred to by Le Roux and Lebaron (2015).

Case studies that will be covered include the relevance of profile and trend aggregation, that both both benefit from the principle of distributional equivalence, and from the information focusing that results from what constitues the principal elements of the analysis, and what constitutes the supplementary elements. (See Murtagh et al., 2015; Murtagh, 2016a, where the analysis is on a set of 12 million Twitter tweets.)

By availing of the remarkable simplicity (Murtagh, 2009, 2016b) of either very high dimensional data, or the data piling of massive data sets, it can be both efficient and effective to carry out our analytics on large data volumes. Case studies in bioinformatics will be covered (Murtagh, 2016d), and the important burgeoning area of both quantitatively and qualitatively evaluating research achievements, and the impact of research.

References

1. N. Keiding and T.A. Louis, "Perils and potentials of self-selected entry to epidemiological studies and surveys", *Journal of the Royal Statistical Society*, Series A, 179 (2), 319–376, 2016. Including F. Murtagh comments.
2. B. Le Roux and F. Lebaron, "Idées-clefs de l'analyse géométrique des données" (Key ideas in the geometric analysis of data), in F. Lebaron and B. Le Roux, Eds., *La Méthodologie de Pierre Bourdieu en Action: Espace Culturel, Espace Social et Analyse des Données*, Dunod, pp. 3–20, 2015.
3. F. Murtagh, "The remarkable simplicity of very high dimensional data: application to model-based clustering", *Journal of Classification*, 26, 249–277, 2009.

4. 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*, 2015, in press. (2016a)

5. F. Murtagh, "Big Data scaling through metric mapping: Correspondence Analysis in very high dimensional spaces", IFCS 2015, International Federation of Classification Societies, Proceedings, submitted. (2016b)

6. F. Murtagh, "Semantic mapping underpinning qualitative and quantitative rankings and ratings, and trending", CLEF 2016, Conference and Labs of the Evaluation Forum, 5-8 September 2016, submitted, 2016. (2016c)

7. F. Murtagh, "Fast scaling followed by linear time clustering of massive data sets in moderate to low dimensions: Exploiting the cloud dual space", in preparation, 2016. (2016d)