Robust regression with compositional covariates in the presence of cellwise contamination

N. Štefelová[1], A. Alfons[2], J. Palarea-Albaladejo[3], P. Filzmoser[4], and K. Hron[1]
[1] Palacký University, Olomouc, Czech Republic; *Nikola.Stefelova@seznam.cz*
[2] Erasmus Universiteit Rotterdam, Rotterdam, The Netherlands
[3] Biomathematics and Statistics Scotland, Edinburgh, UK
[4] Vienna University of Technology, Vienna, Austria

Abstract

Multivariate data are commonly arranged as a rectangular matrix with cases or observations in the rows and variables in the columns. Ordinary robust estimators are designed to deal with entire outlying or contaminated rows, assuming that there is a majority of non-contaminated observations in the data set. However this may not be realistic in many situations where contamination occurs at the cell level. That is, where only a small number of variables is affected per case, but contamination typically propagates throughout many observations. In this case, suppressing entire rows can lead to unacceptable and unnecessary loss of information, particularly in high-dimensional settings (Alqallaf and others, 2009). Additional problems arise when data of compositional nature are involved, because then all the relative information about a certain cell representing a compositional part is contained in ratios of such a part to other parts (Pawlowsky-Glahn and others, 2015). A way to handle this in ordinary regression with real response and explanatory composition was introduced in Hron and others (2012). The robust regression method proposed in this work starts with filtering (i.e. flagging and eliminating) extreme cellwise outliers in compositional parts from the pairwise logratio data matrix. Orthonormal logratio coordinates are optimized based on how often the original parts are involved in outliers. Accordingly, a coordinate system that highlights the role of single compositional parts (Fišerová and Hron, 2011), called recently pivot coordinates, seems to be preferable for this purpose. Finally, the shooting S estimator for regression under cellwise contamination (Öllerer and others, 2016) is adapted to work in coordinate representation. The performance of the procedure is illustrated with real-world biological data.

References

Alqallaf, F., Van Aelst, S., Yohai, V.J., Zamar, R.H., (2009). Propagation of outliers in multivariate data. *The Annals of Statistics 37*(1), pp. 311–331.

Fišerová, E., Hron, K. (2011). On interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences 43*(4), pp. 455–468.

Hron, K., Filzmoser, P., Thompson, K. (2012). Linear regression with compositional explanatory variables. *Journal of Applied Statistics 39*(5), pp. 1115–1128.

Öllerer, V., Alfons, A., Croux, C. (2016). The shooting S-estimator for robust regression. *Computational Statistics 31*(5), pp. 829–844.

Pawlowsky-Glahn, V., Egozcue, J.J., Tolosana-Delgado, R. (2015). *Modeling and analysis of compositional data*. Chichester: Wiley.