HANDLING COMPOSITIONAL VALUES BELOW DETECTION LIMIT IN CHEMICAL DATA

J.A. Martín-Fernández [1], J. Palarea-Albaladejo [2]

[1]*Dept. Computer Science and Applied Mathematics, U. of Girona, Edifici P4 Campus Montilivi, Girona, Spain. (josepantoni.martin@udg.edu)*

[2]*Biomathematics and Statistics Scotland, JCMB, The King's Buildings, Edinburgh, EH9 3JZ, UK.*

Compositional data is frequently collected in many applied fields such as chemistry, nutrition and behaviour sciences. Formally, a compositional vector or simply a composition is defined as a D-dimensional vector $\mathbf{x}=[x_1, x_2,\ldots, x_D]$ such that $x_j>0$, j= 1,…,D, subject to a constant-sum constraint $x_1+x_2+ \ldots +x_D =1$. The log-ratio methodology (e.g., [1]) provides a powerful framework to deal with that type of data.

Below-detection-limit (BDL) values frequently occur when handling multivariate chemical concentrations. Typically, a certain true value cannot be measured or observed due to a low concentration of the corresponding substance or element. In these cases, the unknown value is commonly recorded in the data matrix as either a zero or an annotation such as "$<\varepsilon_{ij}$", where ε_{ij} is the threshold or detection limit of the measuring process applied to element *j* in composition *i*. Note that the true value is unknown but the information about its possible maximum value is available. This type of problem is known as "the rounded zero problem" ([2]) in compositional data analysis.

Many multivariate data analysis techniques require complete data matrices. Hence, there is a demand of sensible imputation strategies for BDL values. We can find several replacement techniques for BDL values in the literature: the "multiplicative replacement" [3], the "modified EM alr-algorithm" [4], and the "robust ilr-algorithm" [5]. All these proposals are coherent with the special nature of compositional data, although each one has its particular virtues. On the other hand, when the goal is the estimation of distributional parameters such as mean, median, or percentiles, the strategy must be different. Bootstrap resampling is an attractive, computationally-intensive approach for estimating population parameters and their associated uncertainties. A "non-detect bootstrap" approach [6] based on combining a replacement technique with ordinary bootstrapping will do the job.

A methodology for handling compositional values below detection limit in chemical data is presented. To illustrate this methodology, a case study involving chemical data is conducted.

Acknowledgement: This research was supported by the Ministerio de Ciencia e Innovación under the project "CODA-RSS" Ref. MTM2009-13272; by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under the project Ref: 2009SGR424; and by the Scottish Government.

References:

[1] Pawlowsky, V. and Buccianti, A. (eds), 2011. *Compositional Data Analysis: Theory and Applications*. Chichester (UK), John Wiley & Sons *(ISBN: 978-0-470-71135-4; 378p.)*.

[2] Martín-Fernández, J.A., Palarea-Albaladejo, J., Olea, R.A., 2011. Dealing with zeros, Ch. 4. In [1], pp. 47-62.

[3] Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V., 2003. Dealing with zeros and missing values in compositional data sets using nonparametric imputation. Mathematical Geology 35 (3), 253-278.

[4] Palarea-Albaladejo, J., Martín-Fernández, J.A., 2008. A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. Computers & Geosciences 34 (8), 902-917.

[5] Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P. and Palarea-Albaladejo, J., 2012. Model-based replacement of rounded zeros in compositional data: classical and robust approach. *Computational Statistics & Data Analysis*. (accepted)

[6] Palarea-Albaladejo, J., Martín-Fernández, J.A. and Olea, R.A., 2011. Non-detect bootstrap method for estimating distributional parameters of compositional samples revisited: a multivariate approach. In: Egozcue, J.J., Tolosana-Delgado, R. and Ortego, M.I. (eds.). Proceedings of the 4th Compositional Data Analysis Workshop, May 10-13, Sant Feliu de Guixols, Spain, 9 p.