A PAIRWISE LOG-RATIO METHOD FOR THE IDENTIFICATION OF BIOMARKERS

J. Walach[1], P. Filzmoser, K. Hron, B. Walczak
Institute of Statistics and Mathematical Methods in Economics, TU Wien
Vienna, AUSTRIA
e-mail: [1]jan.walach@tuwien.ac.at

Abstract

One of the main goals in metabolomics is the identification of biomarkers metabolites which are capable of distinguishing between groups of, e.g., healthy and unhealthy patients. There are various methods for identifying biomarkers in the statistical field. Difficulties arise by facing the so-called size effect, which occurs due to different sample volume or concentration. In that case, the true signal is hidden in the data structure, and it can be revealed only after a special treatment. One possibility is to normalize the data first, other possibilities include certain transformations, see e.g. [1].

Here we propose a method that makes use of the log-ratio approach [2]. We use the elements of the variation matrix, which are defined as the variance of $\log(x_i/x_j)$, for all pairs of variables x_i and x_j. The advantage of log-ratios is that the absolute concentration is irrelevant, which is appropriate in this context. The variation matrix is computed for the joint data, as well as for the single groups separately. A statistic is then constructed, involving all three sources of information. Since the distribution of the statistic is unknown, we use the bootstrap technique; biomarkers are then considered as variables where most of their pairwise log-ratios are significantly different.

The method has been tested on simulated data as well as on real data sets. The simulations have been carried out according to the scheme outlined in [1]. In both the low-dimensional (9 variables) and the high-dimensional (500 variables) situation, the new proposal shows excellent behavior with respect to the true positives, false discovery and false negative rates. These simulations reveal slight advantages over PQN normalization, the method which turned out in [1] as the best among all considered options. The new pairwise log-ratio method has the big advantage that it can easily be robustified against outliers in the data, by simply using a robust estimator of the variance.

References

[1] Filzmoser P., Walczak B. (2014). What can go wrong at the data normalization step for identification of biomarkers? *J. Chromatography A*, Vol. 1362, p. 194.

[2] Pawlowsky-Glahn V., Egozcue R., Tolosana-Delgado J.J. (2015). *Modeling and Analysis of Compositional Data*, Wiley, Chichester, UK.