A robust pairwise log-ratio approach for variable selection and cell-wise outlier diagnostics with focus on metabolomic data

J. Walach[1], P. Filzmoser[1], K. Hron[2], B. Walczak[3]and L. Najdekr[2,4]
[1]Vienna University of Technology, Austria *jan.walach@tuwien.ac.at*
[2]Palacky University, Olomouc, Czech Republic
[3]University of Silesia, Katowice, Poland
[4]University Hospital Olomouc, Czech Republic

Abstract

One of the main goals in metabolomics is the identification of diagnostically important metabolites - biomarkers. The statistical field provides various possibilities for identifying biomarkers. Such methods are ordinarily called variable selection methods. Difficulties arise when facing the so-called "size-effect", which occurs due to different sample volume or concentration. In that case, absolute information is no longer interesting but relative information might provide useful results. Here we propose a method that makes use of the log-ratio approach (Pawlowsky-Glahn and others, 2015). We use the elements of the variation matrix 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. For variable selection a statistic involving all three sources of information is constructed (Walach and others, 2016). The method can be easily robustified against outliers in the data by simply using a robust estimator of the variance.

As a secondary tool of the method, cell-wise outlier diagnostics is performed. In data analysis outlier diagnostics usually refers to an analysis of rows of the data matrix. However, often most data cells in a row are regular, and only few are deviating. Thus, cell-wise outlier diagnostics can bring more insight into the data. The robust weights obtained by robust estimation of the variation matrix are aggregated and displayed in a diagnostics plot, which allows to reveal the outlying cells in the data matrix.

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 Filzmoser and Walczak (2014). The variable selection method shows excellent behaviour with respect to the true positives and false discovery rates.

References

Filzmoser, P., Walczak, B. (2014). What can go wrong at the data normalization step for identification of biomarkers? *Journal of Chromatography A 1362*, 194–205.

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

Walach, J., Filzmoser, P., Hron, K., Walczak, B., Najdekr, L. (2016) *Robust biomarker identification in a two-class problem based on pairwise log-ratios* Manuscript submitted for publication