Assessing cutoff points for biomarker selection in -*omics* technologies

Pietro Franceschi[1,⋆], **Ron Wehrens**[1]

1.Biostatistics and Data Management,
Edmund Mach Foundation – Research and Innovation Centre
Via Edmund Mach 1
38010 San Michele all'Adige (TN)
Italy

⋆Contact author: pietro.franceschi@iasma.it

Keywords: Metabolomics, Cutoff Values, Biomarker Selection.

Data from modern -*omics* technologies provide an holistic representation of the state of biological systems. However, considering the inherent complexity of the datasets, it is necessary to develop filters able to highlight only relevant information to be used for the biological interpretation.

Biomarker identification represents a paradigmatic example of the situation faced in data filtering. Biomarkers are variables (metabolites, proteins, genes, ...) which can be used to characterize specific subgroups in the data; in a two-class setting, for example, the biomarkers are those variables that allow discrimination between the classes.

A *class* tag can be used to distinguish many situations: it can be used to discriminate treated vs. non-treated, to mark different varieties of the same organism, etcetera. Most methods for biomarker identification formulate the problem in a classification setting, selecting as important the variables which give good predictive performance.

In all the biomarker selection strategies, it is necessary to define a cutoff value which identifies the subset of the features to be considered biomarkers, as with the rather arbitrary α level in statistical testing.

The choice of a reasonable and robust cutoff has important practical implications for the development of a biological model. Biomarkers identification, indeed, can be a long and expensive process so it is of paramount importance to be able of focus only on *reliable* biomarkers.

In the majority of cases the optimal cutoff point depends on the specific dataset under study, so there is a definitive interest in developing and comparing general strategies able to identify good cutoff points.

In this contribution, several strategies to face the problem will proposed and evaluated using spiked -*omics* datasets. Among them, particular focus will be on *Higher Criticism* (Donoho; 2008) and on the recently proposed *Stability Based Biomarkers Selection* (Wehrens; 2011).

References

D.Donoho, J.Jin (2008). Higher criticism thresholding: Optimal feature selection when useful features are rare and weak. *PNAS 105(39)*, 14790 – 14795.

R. Wehrens, P. Franceschi, U. Vrhovsek, F. Mattivi (2011). Stability-based biomarker selection. *Analytica Chimica Acta, doi:10.1016/j.aca.2011.01.039.*