Meta-statistics for Biomarker Selection in the Omics Sciences

Wehrens, R.[1,*]; Franceschi, P.[1]

[1]Biostatistics and Data Management, Fondazione Edmund Mach

*Presenting author: **Ron Wehrens** (ron.wehrens@fmach.it)

Background
Biomarker selection, i.e., the definition of which variables are important in statistical regression or discrimination models, is an ever more important topic in the omics sciences. Data from these fields are typically characterized by a low number of samples, but a large number of variables – a meaningful biological interpretation often is only possible when considering the most important variables.

Methods
In this context, statistical tests like the t test will lead to many false positives, while multiple testing corrections tend to lose much power and select only very few variables. In addition, the cutoff value (usually set to a value like 5%) is often chosen in a haphazard way. We present two meta-statistics to tackle the problem of variable selection: higher criticism thresholding [1,2] and stability selection [3,4]. Higher criticism thresholding, applicable in a two-class discrimination setting, is a way to set suitable cutoff levels for significance, based on the data at hand. The underlying mechanism has been described as the "z-score of the p-value" [1]. The current work has extended higher criticism to multivariate methods like PLSDA and the VIP statistics [4]. Stability selection is a novel variable selection method, assessing the stability of biomarker selections under perturbations of the data. The concept is extremely general and robust and can be applied both in regression and discrimination cases: primary selection methods assessed in thie work include PLS and lasso models.

Results
Simulated as well as experimental data show very good results for both stability selection and higher criticism. The experimental data in this study consist of LC-MS metabolomics data of spiked-in apple extracts [5] – such spike-in data are extremely important in assessing the value of biomarker selection methods but are rarely available. Good results are also obtained in other areas of science [1-3]. The advantages of stability selection include a broad applicability (regression, discrimination) and modest computational demands; on the other hand, the number of samples that is required is relatively high. For discrimination problems with fewer than, say, eight samples per class, it is probably better to rely on the higher criticism approach. Both higher criticism and stability selection have been implemented in an R package, BioMark, available from the CRAN repository, and also containing the experimental spike-in data.

[1] Donoho, D. and Jin, J. (2004). Higher criticism for detecting sparse heterogeneous mixtures. *Ann. Statist.*, **32**, 962–994.

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

[3] Meinshausen, N. and Buhlmann, P. (2010). Stability selection. *J. R. Statist. Soc. B*, **72**, 417–473. With discussion.

[4] Wehrens, R., Franceschi, P., Vrhovsek, U., and Mattivi, F. (2011). Stability-based biomarker selection. *Anal. Chim. Acta*, **705**, 15–23.

[5] Franceschi, P., Masuero, D., Vrhovsek, U., Mattivi, F., and Wehrens, R. (2012). A benchmark spike-in data set for biomarker identification in metabolomics. *J. Chemom.*, **26**, 16-24.