An efficient large scale prediction and feature selection approach for microarray and proteomics data

Korbinian Strimmer
(joint work with Verena Zuber and Miika Ahdesmäki)

We propose an effective framework [1] for high-dimensional linear discriminant analysis (LDA) based on three key elements: James-Stein shrinkage for learning prediction rules, feature ranking by correlation- adjusted t-scores (cat scores) [2], and feature selection by thresholding and controlling false non-discovery rates (FNDR). Relative to competing LDA approaches (such as SCRDA) our algorithm is computationally inexpensive and makes practical high-dimensional LDA analysis. Furthermore, we show on experimental data sets and by comparing with the "higher criticism" approach that feature selection by FNDR control is very effective not only for LDA but also for diagonal discriminant analysis. The proposed shrinkage discriminant and variable selection procedure is implemented in the R package sda available from the R repository CRAN (`http://cran.r-project.org/web/packages/sda/`).

Keywords: discrimininant analysis, prediction, "small n, large p", feature selection, false discovery rates, higher criticism.

[1] Ahdesmäki, M. and Strimmer, K. (2009). Feature selection in "omics" prediction problems using cat scores and false non-discovery rate control, arXiv:0903.2003.

[2] Zuber, V. and Strimmer, K. (2009). Gene ranking and biomarker discovery under correlation, arXiv:0902.0751.