MSIAM MSc thesis project proposal, Tracks DS or STAT

Optimal testing of nonparametric hypotheses for Gene Set Enrichment Analysis

Anatoli Juditsky and Bernard Ycart[1]

Among many statistical problems that arise in the analysis of genomic data, that of testing if a given gene set is significantly connected with a genome-wide vector of expression data, is of crucial importance. Many tests exist, and at this point it is still not clear which one gives the best (i.e. biologically most relevant) results. [3]. The *Gene Set Enrichment Analysis – GSEA* aims at comparing a vector of expression data indexed by the set of all genes to the genes in a much smaller given set, and the principal question to be answered is: are the expressions inside the gene set significantly different from the weights in an independent random set of the same size. One approach to answering this question relies upon considering ranks of genes of the gene set with respect to the weights defined by the expression data. Note that if there were no particular relation between the weights and the gene set, these ranks would be considered as random sample without replacement. The above setting then can be reduced to the problem of testing the hypothesis of uniformity of the distribution of independent observations on $[0, 1]$, so that "classical" [4] and "modern" (cf. [1, 2]) nonparametric techniques can be applied.

The objective of this project is to develop and analyse new efficient nonparametric tests of the independence hypothesis for GSEA. We aim at the tests with two desired properties:

- the ability to deal with "weighted ranks", which arise naturally in applications and should allow to develop more powerful tests;
- the tests are to be "immunized" against the non-uniformity of the distributions of a typical gene set to keep test's False Discovery Rate (FDR) at practically acceptable levels.

New tests will be implemented in `R`, and compared to the existing methods: computing time, power against given alternatives and robustness to false detections [5]; also the capacity of the test to detect already known associations will be investigated.

The project will be hosted by the LJK lab, under the supervision of both Anatoli Juditsky and Bernard Ycart. Basic knowledge of hypothesis testing `R` language are requested. The MSc project could be the first step towards a PhD thesis.

References

[1] M. S. Ermakov. Asymptotically minimax criteria for testing composite nonparametric hypotheses. *Problems Inform. Trans.*, 32(2):184–196, 1996.

[2] Y. I. Ingster. Asymptotically minimax testing of the hypothesis of independence and other composite hypotheses. *Theory of Prob. and Appl.*, 31(3):558–569, 1987.

[3] K. Charmpi and B. Ycart. Weighted Kolmogorov-Smirnov testing: an alternative for Gene Set Enrichment Analysis. *Statist. Appl. in Genetics and Molecular Biology*, 14(3):279–295, 2015.

[4] E.L. Lehmann, J. P. Romano. *Testing Statistical Hypotheses (Springer Texts in Statistics)*, Springer, 2005.

[5] B. Ycart, F. Pont, and J. J. Fournié. Curbing false discovery rates in interpretation of genome-wide expression profiles. *J. Biomed. Inform.*, 47:58–61, 2014.

[1]LJK, CNRS UMR 5224, Univ. Grenoble Alpes, France, 04 76 51 49 95 `Anatoli.Juditsky, Bernard.Ycart@imag.fr`