Proposing Genetic risk score based on statistical learning.

Michael Blum, CNRS senior researcher
http://membres-timc.imag.fr/Michael.Blum/
Laboratoire TIMC-IMAG
Faculté de medicine, La Tronche
michael.blum@imag.fr
04 56 52 00 65

A PhD subject can be proposed following the MSc thesis. The proposal pertains to students interested in statistics and data science with applications to health science. A good knowledge of regression approaches (least-squares, regularization) is preferred as well as some experience in programming languages (C, Python, R).

Genetic risk score or polygenic risk scores (PRSs) combine information from multiple genes into a single score for predicting disease risk. Although, each individual gene may have a small predictive power, a score that combines data from multiple genes can be a strong predictor of disease. In the future, public health strategies will use polygenic scores to identify high-risk individuals where disease prevention interventions should be focused (Wray et al. 2013).

Current statistical models for generating genetic score are based on linear or linear mixed models (Dudbridge 2013, Golan and Rosset 2014). Linear models assume that genetic scores are equal to the sum of the effect of each gene. However, genes interact with each other indicating that higher order interaction should be accounted for (Franberg et al. 2015). The objective of the MSc thesis is to propose and implement statistical models that account for interaction between genes. The main difficulty is that the number of genetic markers is of the order of one million making the number of pairwise interactions of the order of one thousand billion. To handle huge number of possible interactions, group-lasso regularization is an example of well-suited learning approach (Lim and Hastie 2014).

Prediction scores of different learning models will be compared for different diseases. Comparisons will be based on the GERA database that contains data for around 100,000 individuals. Individuals in the database consists of “control” individuals and of sick individuals for a variety of diseases that affect people in adulthood including depression, insomnia, and diabetes.

References

Dudbridge, Frank. "Power and predictive accuracy of polygenic risk scores."*PLoS Genet* 9.3 (2013): e1003348.
http://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003348

Frånberg, Mattias, et al. "Discovering Genetic Interactions in Large-Scale Association Studies by Stage-wise Likelihood Ratio Tests." *PLoS Genet* 11.9 (2015): e1005502.
dx.plos.org/10.1371/journal.pgen.1005502

Lim, Michael, and Trevor Hastie. "Learning interactions via hierarchical group-lasso regularization." *Journal of Computational and Graphical Statistics* just-accepted (2014): 00-00.
http://www.web.stanford.edu/~hastie/Papers/glinternet.pdf

Golan, David, and Saharon Rosset. "Effective genetic-risk prediction using mixed models." *The American Journal of Human Genetics* 95.4 (2014): 383-393.
http://europepmc.org/articles/pmc4185122

Wray, Naomi R., et al. "Pitfalls of predicting complex traits from SNPs." *Nature Reviews Genetics* 14.7 (2013): 507-515.
http://europepmc.org/articles/pmc4096801

Information about the GERA dataset
https://www.nia.nih.gov/research/blog/2014/06/genetics-data-available-secondary-analysis