Augmented prediction in high-dimensional omic applications

Mar Rodríguez Girondo[1]

[1]M.Rodriguez_Girondo@lumc.nl, Department of Medical Statistics, Leiden University Medical Center

During the past decade, much attention has been devoted to accommodate single high-dimensional sources of molecular data (omics) in the calibration of prediction models for health traits. Genomic and transcriptomics data have been established as widely used omic predictors of health outcomes, alone or in combination with (low-dimensional) clinical covariates. Nowadays, new omic sources, such as proteomics, metabolomics, and glycomics are emerging as potentially interesting molecular sources for prediction of health-related traits. As a result, it is increasingly common to collect several omic measurements in the same set of individuals, and hence, new statistical challenges are emerging nowadays, namely, how to combine all these new information and to quantify the additional value of new molecular sources over previously established ones. Our motivating example illustrates these difficulties. We consider the Dietary, Lifestyle, and Genetic determinants of Obesity and Metabolic syndrome (DILGOM) study, sampled from the Helsinki area, in Finland. Data on serum metabonomes, genome-wide profiles of genetic and transcriptional variation from blood leukocytes are available, jointly with a large number of clinical and demographic factors. We are interested in investigating the role of each of the omic sources in the prediction of BMI and obesity and their respective additional predictive value with respect to each other. We propose a two-step procedure based on sequential double cross-validation prediction and regularized regression models, i.e., we consider the problem of combination of omic predictors in an `asymmetric' way by sequentially assessing the augmented predictive ability of omic sources with respect to a given outcome of interest. Namely, we propose several performance indices to summarize the relation between the omic sources under study and a permutation test to formally assess the augmented predictive value of a second omic set of predictors over a primary omic source. The abstract must be written in English and it does have to not exceed one page**.** The contents will be the sole responsibility of the authors.

Keywords: Augmented prediction; double cross-validation; regularized regression.