A Multivariate Framework for Variable Selection and Identification of Biomarkers in High-Dimensional Omics Data

Verena Zuber
Dissertation

Abstract

In this thesis, we address the identification of biomarkers in high-dimensional omics data. The identification of valid biomarkers is especially relevant for personalized medicine that depends on accurate prediction rules. Moreover, biomarkers elucidate the provenance of disease, or molecular changes related to disease. From a statistical point of view the identification of biomarkers is best cast as variable selection. In particular, we refer to variables as the molecular attributes under investigation, e.g. genes, genetic variation, or metabolites; and we refer to observations as the specific samples whose attributes we investigate, e.g. patients and controls. Variable selection in high-dimensional omics data is a complicated challenge due to the characteristic structure of omics data. For one, omics data is high-dimensional, comprising cellular information in unprecedented details. Moreover, there is an intricate correlation structure among the variables due to e.g internal cellular regulation, or external, latent factors. Variable selection for uncorrelated data is well established. In contrast, there is no consensus on how to approach variable selection under correlation.

Here, we introduce a multivariate framework for variable selection that explicitly accounts for the correlation among markers. In particular, we present two novel quantities for variable importance: the correlation-adjusted t (CAT) score for classification, and the correlation-adjusted (marginal) correlation (CAR) score for regression. The CAT score is defined as the Mahalanobis-decorrelated t-score vector, and the CAR score as the Mahalanobis-decorrelated correlation between the predictor variables and the outcome. We derive the CAT and CAR score from a predictive point of view in linear discriminant analysis and regression; both quantities assess the weight of a decorrelated and standardized variable on the prediction rule. Furthermore, we discuss properties of both scores and relations to established quantities. Above all, the CAT score decomposes Hotellingâ€™s T 2 and the CAR score the proportion of variance explained. Notably, the decomposition of total variance into explained and unexplained variance in the linear model can be rewritten in terms of CAR scores.

To render our approach applicable on high-dimensional omics data we devise an efficient algorithm for shrinkage estimates of the CAT and CAR score. Subsequently, we conduct extensive simulation studies to investigate the performance of our novel approaches in ranking and prediction under correlation. Here, CAT and CAR scores consistently improve over marginal approaches in terms of more true positives selected and a lower model error. Finally, we illustrate the application of CAT and CAR score on real omics data. In particular, we analyze genomics, transcriptomics, and metabolomics data. We ascertain that CAT and CAR score are competitive or outperform state of the art techniques in terms of true positives detected and prediction error.