Targeted learning for high-dimensional variable importance

Alan Hubbard*

hubbard@stat.berkeley.edu;hubbard@berkeley.edu

Exploratory analysis of high dimensional biological data has received much attention since the explosion of high-throughput technology allows simultaneous screening of thousands of characteristics (genetic sequencing, genomics, metabolomics, proteomics, microbiomics, etc.). In this talk, we focus on 1) how to derive estimation of independent associations (variable importance measures) in the context of many competing causes in a semiparametric model, and 2) ideas for robustifying small-sample inference when data adaptive techniques are used. This paper focuses on a Targeted Learning variable importance approach, combining motivation for estimands from causal inference and potentially automated forms of estimation of the data-generating distribution using ensemble machine learning methods (Superlearning). We apply these approaches to high dimensional data sets of relatively modest sample size. Specifically, the analysis is faced with not just a large number of comparisons, but also trying to tease out of association of biomarkers and a phenotype, apart from confounds such as sample characteristics and other biomarkers. The methodology combines existing targeted maximum likelihood learning methods (TMLE) and with a simple generalization of commonly used empirical Bayes approaches to boost performance of small sample inference. Specifically, we propose using TMLE for estimating variable importance measures along with a general adaptation of the commonly used LIMMA approach based upon the so-called estimator influence curve. The result is a machine-based approach that can estimate independent associations in high dimensional data, but offers some protection against the unreliability of small-sample inference. Time-permitting, we also discuss adaptive methods for variable reduction to avoid some of the pitfalls of many multiple comparisons.

This is joint work with Nima Hejazi and Wilson Cai.

*School of Public Health, University of California at Berkeley, Division of Biostatistics, 113B Haviland Hall, Berkeley, CA 94720-7358, USA