Targeted Learning for Variable Importance

Alan Hubbard, Ph.D.
Professor and Division Head
Department of Biostatistics
University of California Berkeley
Berkeley, CA

Abstract:
Common question addressed in exploratory data analysis is: which variables are important? Commonly, variable importance is an exercise in ranking variables with regard to the "importance" in fitting a prediction function. As a result, some prediction procedures have built in or implied variable importance output, such as coefficients in regression approaches (glm, stepwise, lasso, etc) or as changes in the prediction accuracy by backwards removal (e.g., random forest). In fact, there is a commonly stated desire to have both the best prediction function, and also one that yields an "interpretable" model fit, that is, something that is not a black box. We advocate for a different approach that uses both machine learning (the black box) and the Targeted Learning framework to define interpretable variable importance measures. We discuss estimates and associated inference that can be used to rank the importance of predictor variables. The variable importance parameters suggested are motivated by causal (intervention-based), and can return estimates on equivalent scales. We discuss data-adaptive parameter versions of variable importance parameters, as well as show work applied to health care (acute trauma) data.

Biography:
Alan Hubbard, Professor of Biostatistics, Univ. of California, Berkeley, is the Principal Investigator of a study of statistical methods related to patient-centered outcomes research among acute trauma patients (PCORI), head of the computational biology Core E of the SuperFund Center at UC Berkeley (NIH/EPA), as well a consulting statistician on several federally and foundation projects. He has published over 200 articles, including numerous peer-reviewed articles related to diarrheal disease. He has worked as well on projects ranging from molecular biology of aging, epidemiology, and infectious disease modeling, but most all of his work has focused on semi-parametric estimation in high-dimensional data. His current methods-research focuses on precision medicine, variable importance, statistical inference for data-adaptive parameters, and statistical software implementing targeted learning methods. Currently working in several areas of applied research, including early childhood development in developing countries, patient outcomes from acute trauma, environmental genomics and comparative effectiveness research in diabetes care.