UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

BIOSTATISTICS SEMINAR

Jennifer A. Sinnott
Research Associate
Biostatistics and Epidemiology
Harvard School of Public Health

Statistical Learning Methods for Risk Prediction

A pressing goal in cancer research is to develop better models for predicting prognosis by identifying novel biomarkers that can enhance models using established clinical factors alone. When many potential biomarkers are available, such as in studies relating tumor gene expression to cancer survival, many techniques have been developed to identify and incorporate important genes or pathways of genes in order to improve risk prediction. In this talk, I will discuss various testing and prediction methods with high dimensional biomarkers using kernel machine regression and regularized estimation methods. One focus will be given to building risk prediction models using genes which can be naturally grouped into pathways. We discuss kernel machine regression methods for capturing and estimating the potentially complex pathway effects. Kernel-based methods are becoming more frequently used in medical research, in part because different kernels allow for different assumptions about the relationship between genes in the pathway and outcome; however, it is not generally known in advance which kernel is most appropriate for a data set. We discuss omnibus testing methods to combine information across kernels, and kernel selection methods for optimal prediction. When multiple pathways are under consideration, we may use the multiple kernel learning framework for combining information across pathways. A second focus will be given to the prediction of survival when a regularization method, such as the Lasso, Adaptive Lasso, or Adaptive Elastic Net, is used to simultaneously select important genes and fit a regression model, which can then be used to predict survival for new patients. Providing accurate inference for this survival prediction can be challenging; specifically, we demonstrate that existing formulas for the variance of the coefficients based on asymptotic results tend to underestimate the variability for some coefficients, while estimates using standard resampling such as the bootstrap tend to overestimate it, which can both lead to inaccurate variance estimation for predicted survival. We propose an adaptation of a resampling approach that brings the estimated error in line with the truth. Our proposed adaptation uses an ensemble-type approach to look across the coefficient vectors estimated using resampling and allow them to vote on whether each coefficient should be 0. We find that our resampling with voting method is most robust across simulation settings, with coverage typically lying between the asymptotic-based method (which can undercover) and the bootstrap method (which tends to overcover).

Thursday, February 26, 2015 3:30-4:30 PM

133 Rosenau Hall