

1) Modes and Clustering for Time-Warped Gene Expression Profile Data and 2) Random Forest-based Pre-validation Applied to Tissue Microarray Data

Xueli Liu, Ph.D.
University of California, Los Angeles

Abstract:

This talk is comprised of two parts.

In the first part, I will talk briefly about my Ph.D. thesis work. We propose a functional convex synchronization model, under the premise that each observed curve is the realization of a stochastic process. Monotonicity constraints on time evolution provide the motivation for a functional convex calculus with the goal of obtaining sample statistics such as a functional mean. We derive a functional limit theorem and asymptotic confidence intervals for functional convex means. This nonparametric time-synchronized algorithm is also combined with an iterative mean updating technique to find an overall representation that corresponds to a mode of a sample of gene expression profiles, viewed as a random sample in function space.

In the second part, I will talk about novel statistical methods for the analysis of tissue microarray data. Tissue microarrays (TMAs) represent a high throughput tool for studying protein expression patterns in tissue specimens. In performing TMA analysis, the tissue is immunohistochemically stained and scored by a pathologist based on tumor marker staining scores. It is standard practice to select a single staining cutoff that stratifies the population based on an endpoint of interest. However, if the dichotomized staining score is included in a Cox model that uses the same outcome that was used to dichotomize the staining data, the significance of the biomarkers may be overstated. A random forest (L. Breiman 2001) predictor is an ensemble of individual classification tree predictors. Random forests are a state-of-the-art supervised learning method, which is well-suited for data with many covariates but relatively few observations. Here we will explore the use of these predictors for analyzing DNA and tissue microarray data. We introduce a new method (random forest prevalidation) which circumvent this problem. The idea is to summarize all staining scores into a single scalar M which can be used as covariate in a Cox regression model. Specifically, we will use random forest predictors (Breiman 2001) to arrive at an out-of-bag estimates of a hazard score (deviance residual of an intercept only Cox regression model). This procedure is similar in spirit to the pre-validation procedure studied by Tibshirani and Efron (2002), but we show with simulations that it avoids a leakage of degree of freedom. To understand the nature of the dependence of M on the most important covariates we use partial dependence plots and regression trees involving the most important covariates. We demonstrate the use of this method to assess the prognostic significance of eight biomarkers for predicting survival in patients with renal cell carcinoma. Our proposed method avoids problems associated with multi-collinearity and over-fitting. We also carry out a cross-validation scheme to compare the predictive power of different prognostic models.