Yale SCHOOL OF PUBLIC HEALTH
Biostatistics

SEMINAR

Feature Screening for Ultrahigh Dimensional Data

Runze Li, Ph.D.
Professor
Department of Statistics
The Pennsylvania State University at University Park

ABSTRACT

This talk is concerned with screening features in ultrahigh dimensional data analysis, which has become increasingly important in diverse scientific fields. I will first introduce a sure independence screening procedure based on the distance correlation (DC-SIS, for short). The DC-SIS can be implemented as easily as the sure independence screening procedure based on the Pearson correlation (SIS, for short) proposed by Fan and Lv (2008). However, the DC-SIS can significantly improve the SIS. Fan and Lv (2008) established the sure screening property for the SIS based on linear models, but the sure screening property is valid for the DC-SIS under more general settings including linear models. Furthermore, the implementation of the DC-SIS does not require model specification (e.g., linear model or generalized linear model) for responses or predictors. This is a very appealing property in ultrahigh dimensional data analysis. Moreover, the DC-SIS can be used directly to screen grouped predictor variables and for multivariate response variables. We establish the sure screening property for the DC-SIS, and conduct simulations to examine its finite sample performance. Numerical comparison indicates that the DC-SIS performs much better than the SIS in various models. We also illustrate the DC-SIS through a real data example. If time is permitted, I will introduce some newly developed model free screening procedure for categorical high-dimensional data.

4:15 p.m. Tuesday, April 2, 2013

LEPH 115, 60 College Street