STATISTICS COLLOQUIUM

TUESDAY, APRIL 23, 2013
TALK: 4:15 PM — SCIENCE CENTER RM. 705
RECEPTION: 3:50 PM — SCIENCE CENTER, 7TH FLOOR

“Partition Models for Variable Selection and Interaction Detection”

Bo Jiang
Department of Statistics
Harvard University

ABSTRACT

Variable selection methods play important roles in modeling high-dimensional data and are keys to data-driven scientific discoveries. Traditional variable selection methods are derived from a forward modeling angle on the conditional distribution of the response given predictors.

Instead of building a predictive model of the response given combinations of predictors, we start by modeling the conditional distribution of predictors given the response. This inverse modeling perspective motivates us to propose a stepwise procedure that is effective and computationally efficient in detecting interaction with little assumptions on its parametric form. The proposed procedure is able to detect pairwise interactions among p predictors with a computational time of $O(p)$ instead of $O(p^2)$ under moderate conditions. Consistency of the proposed procedure in variable selection under a diverging number of predictors and sample size is established. Its excellent empirical performance in comparison with some existing methods is demonstrated through simulation studies as well as applications in high throughput biological studies.

We will further explore a Bayesian generalization of the proposed procedure --- Bayesian partition model that combines forward and inverse modeling perspectives for selecting both response and predictor variables. In particular, we will discuss the application of partition models in two biological problems: identifying transcription factor subclasses on protein binding microarray, which can be viewed as a high-dimensional clustering with variable selection, and detecting pleiotropic and epistasis effects in expression quantitative trait loci (eQTLs) study, which can be viewed as a multivariate regression with variable selection on both responses and predictors, including also multi-way interactions among correlated predictors.