DEPARTMENT OF STATISTICS SEMINAR SERIES

SIDNEY SMITH HALL, ROOM SS1083

THURSDAY, 18 OCTOBER 2012 AT 3:30PM

Statistical Significance of Clustering for High Dimensional Data

Yufeng Liu

Department of Statistics and Operations Research
Carolina Center for Genome Sciences
The University of North Carolina at Chapel Hill

Clustering methods provide a powerful tool for the exploratory analysis of high dimensional datasets, such as gene expression microarray data. A fundamental statistical issue in clustering is which clusters are "really there," as opposed to being artifacts of the natural sampling variation. In this talk, I will present Statistical Significance of Clustering (SigClust) as a cluster evaluation tool. In particular, we define a cluster as data coming from a single Gaussian distribution and formulate the problem of assessing statistical significance of clustering as a testing procedure. Under this hypothesis testing framework, the cornerstone of our SigClust analysis is accurate estimation of those eigenvalues of the covariance matrix of the null multivariate Gaussian distribution. In this talk, we propose a likelihood based soft thresholding approach for the estimation of the covariance matrix eigenvalues. Our theoretical work and simulation studies show that our proposed SigClust procedure works remarkably well. Applications to some cancer microarray data examples demonstrate the usefulness of SigClust.

Light refreshments will be served at 3:10 p.m.