MICHIGAN STATE UNIVERSITY
Department of Statistics and Probability

COLLOQUIUM

Jiashun Jin

Department of Statistics
Carnegie Mellon University

Spectral Clustering

Tuesday, August 28, 2014
10:20 a.m. - 11:10 am
Refreshments 10:00 am
C405 Wells Hall

Abstract

Consider a two-class clustering problem where we have (X_i, Y_i), $1 \le i \le n$, from two possible classes. The X_i's are $p \times 1$ vectors that are observable, and $Y_i \in \{-1, 1\}$ are class labels which are unknown to us and it is of interest to estimate them.

We propose the following approach to spectral clustering:

1. We use Kolmogorov-Smirnov statistic to assess the importance of the features (i.e., genes).
2. Based on the p-values, we perform a feature selection, where the threshold is determined by the recent idea of Higher Criticism Thresholding (HCT). HCT was proposed before for classification, and we must modify it carefully for clustering.
3. Based on all retained features, we obtain the leading eigenvector of the so-called *dual covariance matrix*, and predict the class labels by the signs of the coordinates of this eigenvector.

We reveal a surprising connection between the HCT and the so-called Signal Noise Ratio (SNR) associated with the post-screening dual empirical covariance matrix. We apply the approach to several gene microarray data sets, where it gives much more satisfactory results than existing clustering methods.

To request an interpreter or other accommodations for people with disabilities, please call the Department of Statistics and Probability at 517-355-9589.